UNIVERSITI MALAYA

FAKULTI PERUBATAN
Faculty of Medicine

MODELLING PUBLIC HEALTH DATA

MQB 7046

SEMESTER 2 2023/2024

ASSIGNMENT 4

MACHINE LEARNING MODEL

NO.	GROUP MEMBER	STUDENT ID
1	CHONG CHEAN TAT	22108105
2	AHMAD ZAKIY BIN TUMADI	22111484
3	MUHAMMAD SOLIHIN BIN REZALI	22110526

Table of Contents

1.0 INTRODUCTION

1.1 Background

Liver disease is a broad term encompassing various conditions that can affect the liver. The liver plays a crucial role in the body's digestive system and overall health, so it is vital to understand the potential causes and risk factors for liver disease. By gaining a deeper understanding of what can lead to liver disease, individuals can take proactive steps to reduce their risk and maintain liver health, while health organizations can educate and inform the public about prevention and early detection. Liver diseases are categorized into several types, each with its causes and risk factors. The total number of cases of Chronic Liver Disease (covering all stages of disease severity) is approximately 1.5 billion globally (Moon et al., 2020). The leading causes of prevalent disease are non-alcoholic fatty liver disease (NAFLD) (59%), Hepatitis B Virus (HBV) (29%), Hepatitis C Virus (HCV) (9%), and alcoholic liver disease (ALD) (2%). Other liver conditions, such as primary biliary cholangitis, primary sclerosing cholangitis, alpha-1-antitrypsin deficiency, Wilson's disease, and autoimmune hepatitis, comprise 1% of the cases (Asrani et al., 2019). Metabolic dysfunction-associated fatty liver disease (MAFLD) is a proposed new name for NAFLD. This modification is designed to recognize the fact that multiple causes of chronic liver disease can frequently co-occur in a patient instead of solely attributing the disease to its underlying cause. The metabolic syndrome consists of a collection of cardio-metabolic dysfunctions, such as elevated fasting plasma glucose and waist circumference, along with higher blood pressure and triglyceride levels, while lower HDL cholesterol levels are observed (Alberti et al., 2009). This updated terminology is anticipated to significantly enhance patients' identification of this condition and connection to healthcare and treatment options (Tan et al., 2021).

Liver diseases are rising in Malaysia, with an estimated prevalence of 20-40% based on previous studies (Chan et al., 2022). MAFLD is reported as the most prevalent. In 2023, its prevalence reached 28.2%, a higher rate than in 2013 (22.7%) (IPH, 2024). Alongside MAFLD, HBV and HCV are also frequent contributors to liver disease in Malaysians, collectively accounting for the highest number of liver-related deaths in the country. Metabolic syndrome is more prevalent in Malaysia (35.9%) compared to the global average of 31.4%, as well as other countries in the region such as Indonesia (28.4%) and Singapore (26.9%) (Noubiap et al., 2022).

Untreated liver diseases can advance to conditions such as cirrhosis and hepatocellular carcinoma (HCC), the most prevalent form of primary liver cancer and a leading cause of liver cancer-related fatalities worldwide (McGlynn et al., 2020). Some other common causes of liver disease include excessive alcohol consumption, autoimmune disorders, and certain medications or toxins. In addition to these factors, genetic predisposition, obesity, and a poor diet high in processed foods and sugar can also contribute to the development of liver disease.

1.2 Literature Search

Age

Non-alcoholic fatty liver disease (NAFLD) affects mainly the middle-aged and the elderly, given that the risk factors for its development tend to increase in prevalence with advancing age (Wang et al., 2013). Age-related alterations significantly impact hepatic function. Hepatic blood flow and volume demonstrably decline, with reductions of up to 33% and 25%, respectively, reported between young adulthood and later life. Furthermore, the aging liver exhibits morphological changes, including a decrease in hepatocyte number and an increase in size and ploidy. Additionally, a reduction in mitochondrial number suggests a potential decline in oxidative respiration (Schmucker et al., 1998). The prevalence of metabolic syndrome demonstrably increases with advancing age (Reynolds et al., 2009). Studies have shown a strong correlation between metabolic syndrome and NAFLD, suggesting a potential bidirectional relationship between the two conditions.

Gender

The understanding of the reasons behind gender disparities in the incidence, progression, and outcomes of prevalent liver diseases remains incomplete. Potential contributing factors encompass the influence of sex hormones on metabolic and oxidative pathways, the observed sex-based variations in gene expression following hepatic injury, and immunological regulatory differences between men and women. Notably, women exhibit a higher prevalence of acute liver failure, autoimmune hepatitis, benign liver lesions, primary biliary cirrhosis, and toxin-induced hepatotoxicity. Conversely, men are more likely to develop malignant liver tumors, primary sclerosing cholangitis, and viral hepatitis. Interestingly, women with hepatitis C virus infection exhibit a lower rate of decompensated cirrhosis, and no significant difference in survival is observed for alcohol-related liver disease compared to men. Furthermore, women diagnosed with hepatocellular carcinoma demonstrate improved survival rates. Overall, mortality from chronic liver disease and cirrhosis is demonstrably two-fold higher in men compared to women (Guy et al., 2013).

Ethnicity

Ethnicity plays a significant role in the prevalence and outcomes of liver diseases. Studies show that diverse ethnic groups have different susceptibilities to liver diseases and distinct etiology patterns. For instance, in Malaysia, there is a higher proportion of ethnic Malays and Indians among patients with non-alcoholic fatty liver disease compared to Chinese, indicating possible disparities in NAFLD prevalence within the Malaysian population based on ethnicity (Lim et al., 2022). Additionally, another research has suggested that Hispanics residing in the United States experience an increasing prevalence of liver disease and a greater risk of mortality when compared to Caucasians (Mathur et al., 2009). It is essential to comprehend these ethnic discrepancies to create successful strategies for preventing, diagnosing, and treating liver diseases while considering the distinct requirements of diverse ethnic groups.

Liver Enzymes

Liver function tests (LFTs) are crucial in diagnosing liver disorders, tracking liver well-being in different situations, and forecasting results in various medical contexts. Markers such as AST, ALT, and LDH often show increased levels in liver diseases and are, therefore, significant indicators. LFTs are also commonly used to screen for liver disease, providing essential diagnostic information (Ahmed et al., 2018). NAFLD is a common reason for abnormal results of these tests globally (Wong et al., 2011).

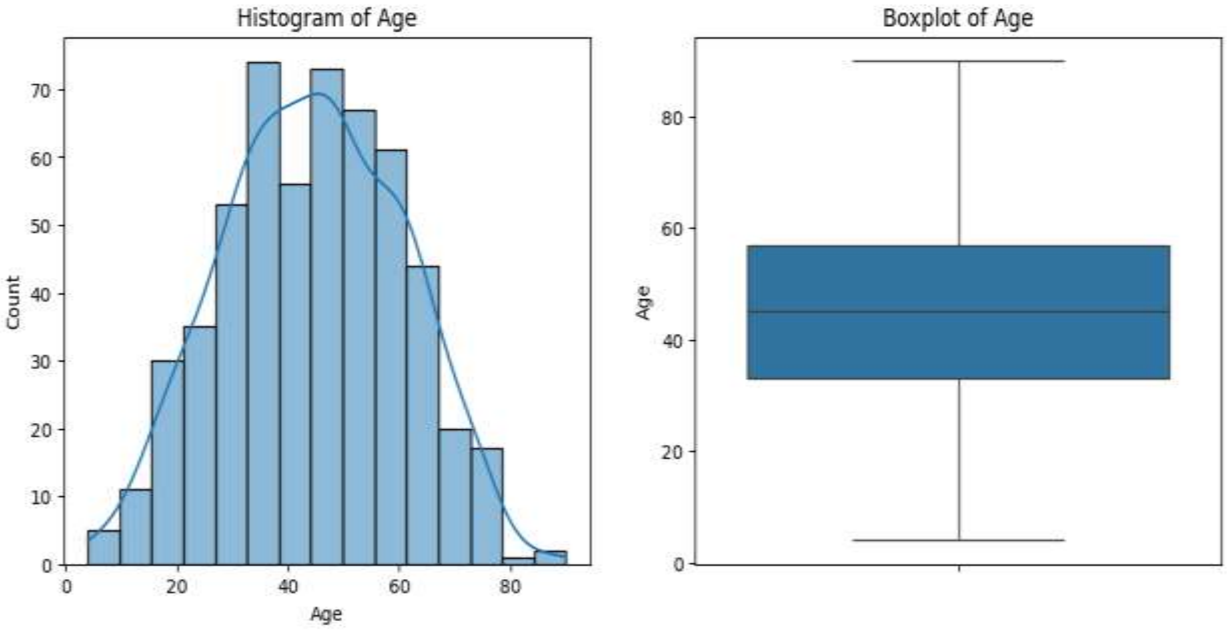
Body Mass Index

Adiposity exerts a demonstrably negative impact on various metabolic functions within the liver. It is strongly associated with the development of steatosis and inflammation characteristic of NAFLD, and it demonstrably promotes the progression of several other liver pathologies, including hepatitis C and alcoholic liver disease (Corey et al., 2014). The prevalence of both NAFLD and its more severe form, NASH, exhibits a direct correlation with increasing body mass index (BMI), with a particularly dramatic rise observed in individuals classified as obese. Furthermore, elevated BMI and the presence of diabetes mellitus are both independently linked to advanced fibrosis in patients diagnosed with NASH (Jimba et al., 2005). The Hepatitis C Antiviral Long-term Treatment Against Cirrhosis (HCV ATLAS) trial conducted a longitudinal study on individuals with chronic hepatitis C who presented with cirrhosis or advanced fibrosis. Significantly, patients who experienced a weight gain exceeding 5% during the first year exhibited a 35% increased risk for a composite outcome encompassing death, hepatic decompensation, and a worsening of fibrosis when compared to those whose weight remained stable (Everhart et al., 2009).

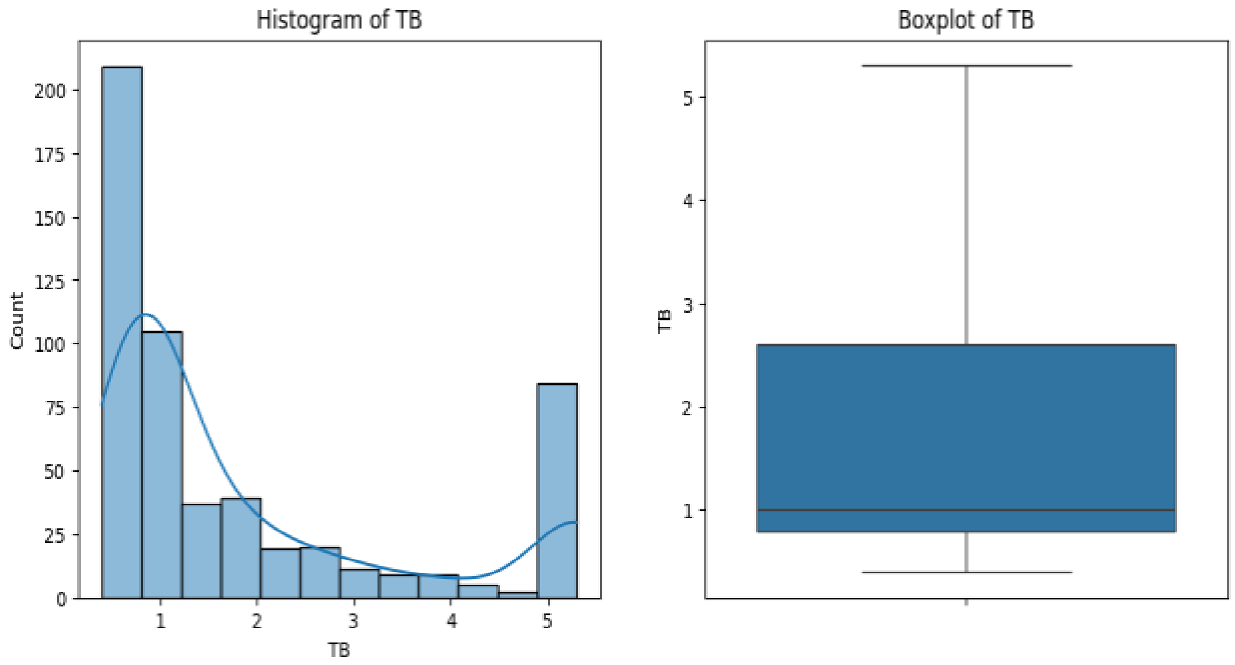
1.3 Descriptive statistics of variables

Descriptive statistics summarize a dataset's distribution's central tendencies, dispersion, and shape. These statistics help in understanding the underlying patterns and characteristics of the data before proceeding with more complex analyses. The dataset used in this study contains 12 features and one target variable related to liver disease. Below is an interpretation of the descriptive statistics for these variables.

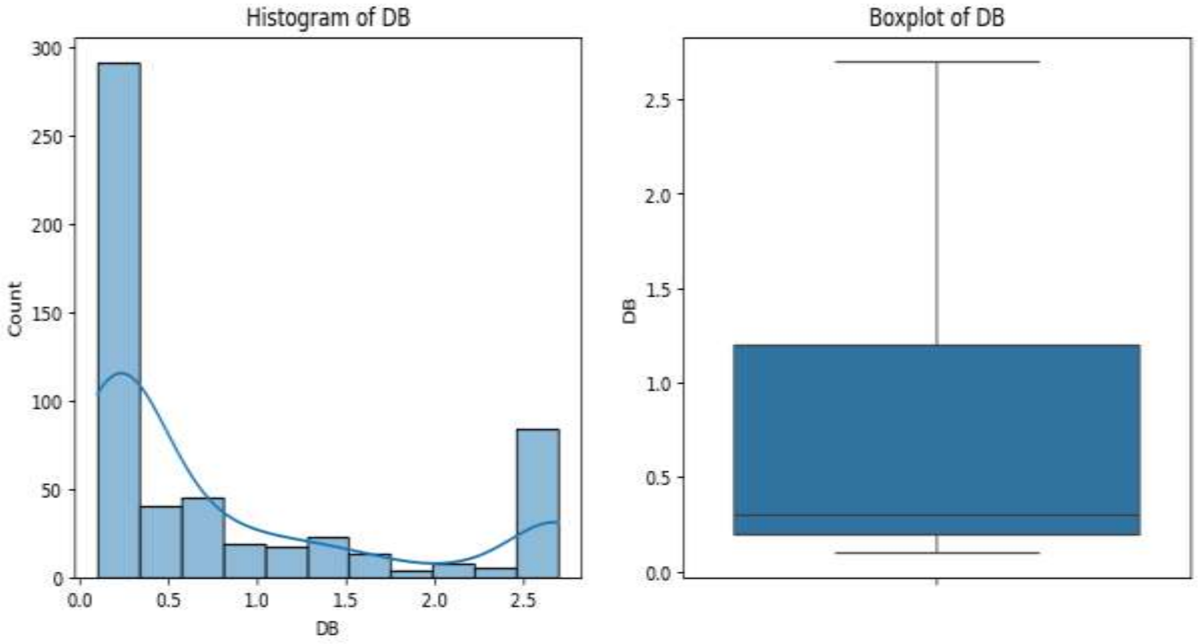
Age: Patients' ages range from 4 to 90 years, with an average (mean) age of approximately 44.5 years. The standard deviation of 16 years indicates a considerable spread in the age distribution, suggesting that the dataset includes both younger and older individuals. The median age, the 50th percentile, is 45 years, indicating a symmetrical distribution around the mean.



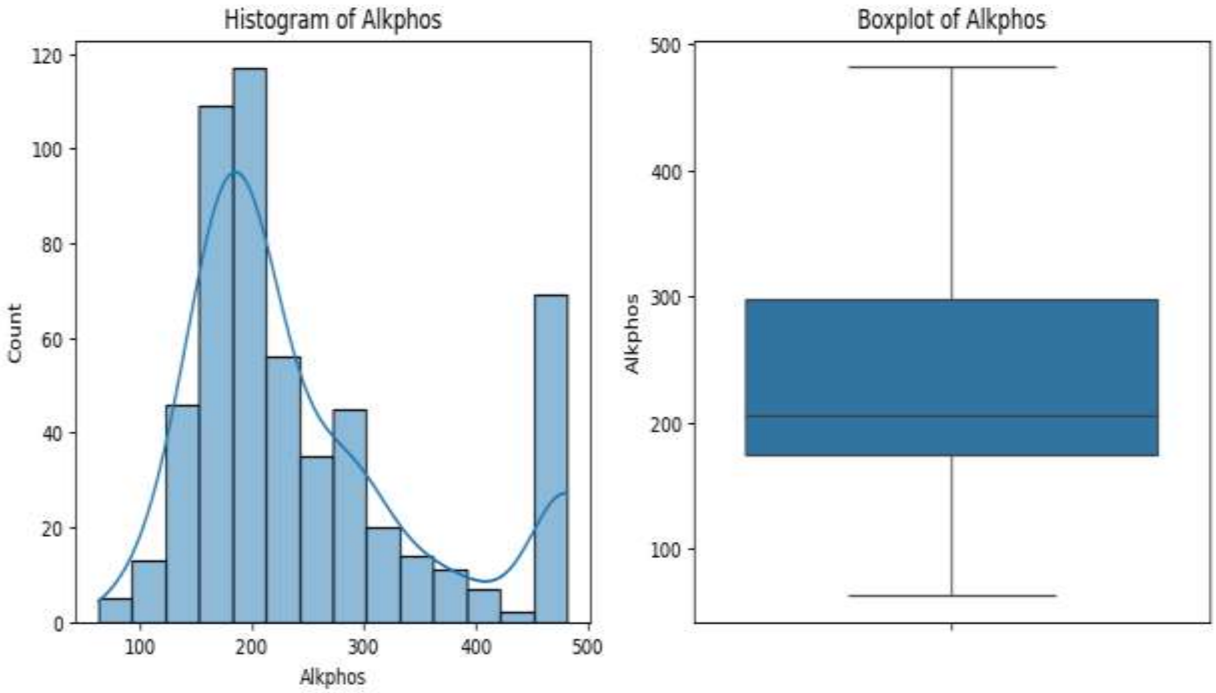
Total Bilirubin (TB): The TB levels vary from 0.4 to 5.3, averaging about 1.91. A higher TB level is often associated with liver dysfunction or damage. The standard deviation of 1.65 suggests substantial variability in TB levels among the patients, which is expected as the dataset includes both healthy individuals and those with liver disease.



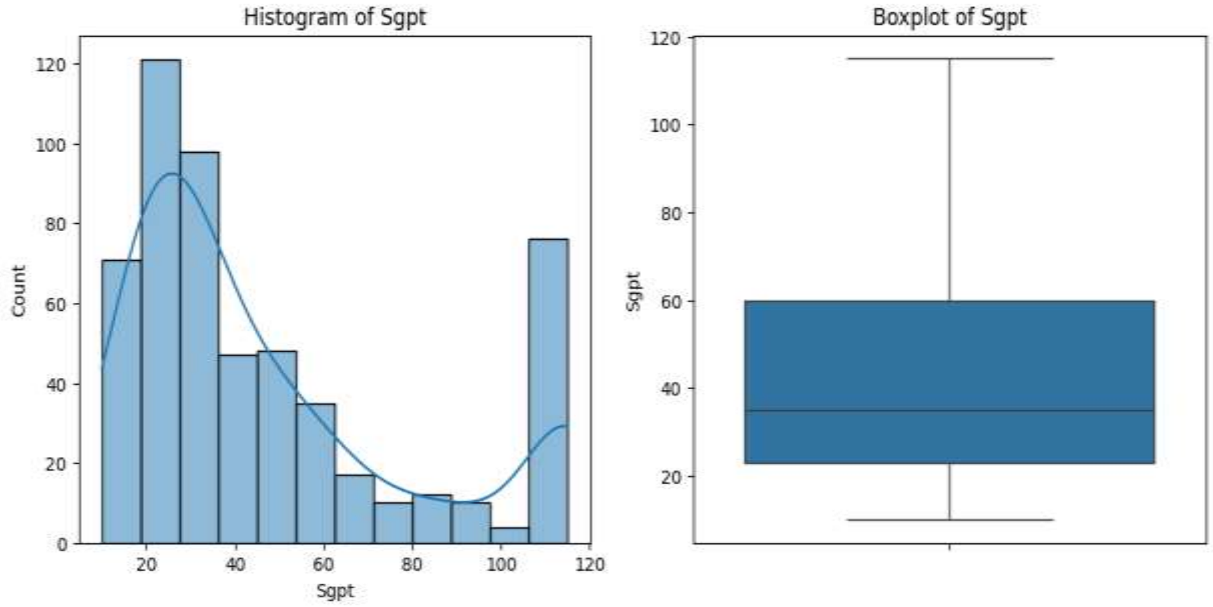
Direct Bilirubin (DB): DB levels range from 0.1 to 2.7, with a mean of around 0.84. The standard deviation is 0.92, indicating variability similar to TB levels. Elevated DB levels can indicate specific liver conditions such as biliary obstruction or hepatitis.



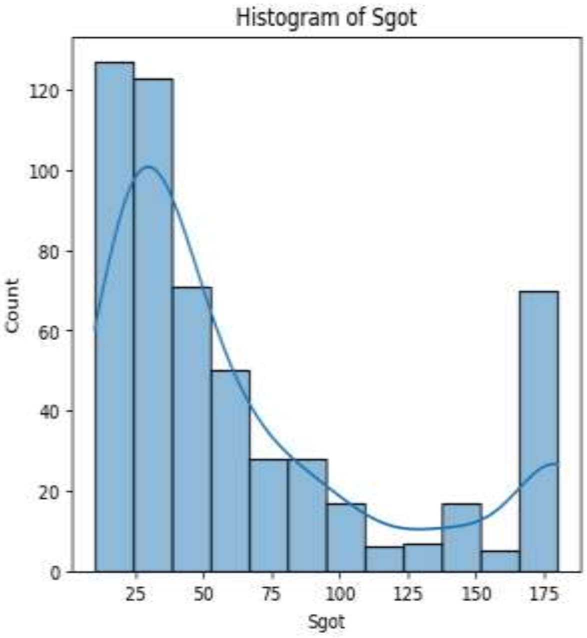
Alkaline Phosphatase (ALP): ALP values range from 63 to 482, with a mean of approximately 248.9. The high standard deviation of 108.9 suggests a wide range of values, reflecting the patients' different liver and bile duct conditions.

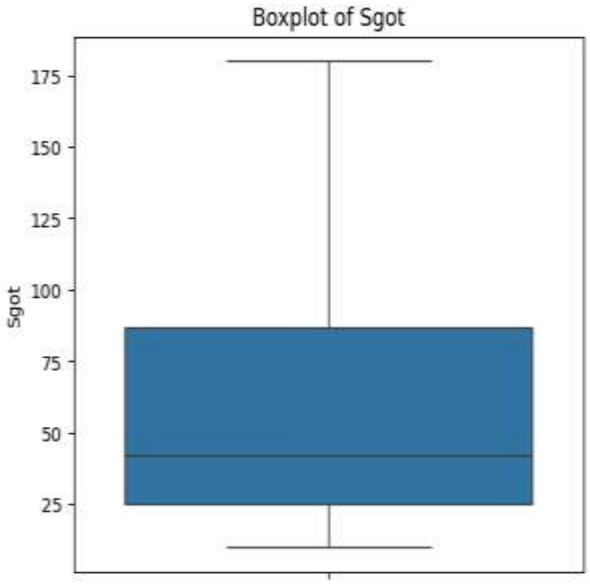


Alanine Aminotransferase (ALT): ALT levels, an enzyme marker for liver function, range from 10 to 115, with a mean of about 47.6. The standard deviation of 32.8 indicates that while many patients have normal levels, there are significant outliers with much higher values, typically indicating liver damage.

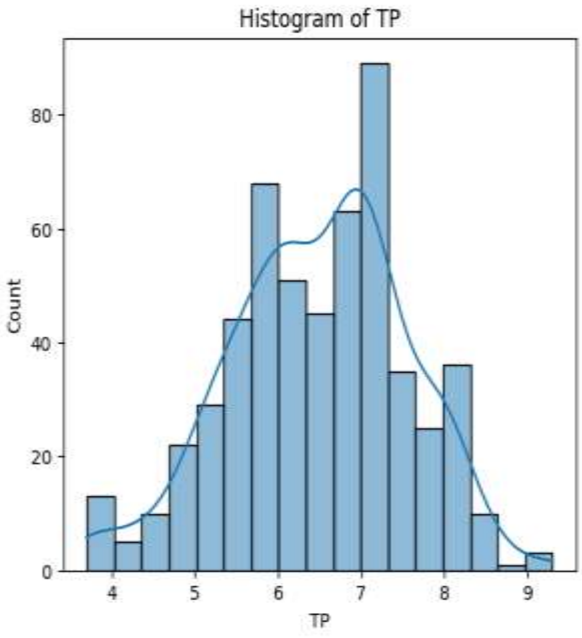


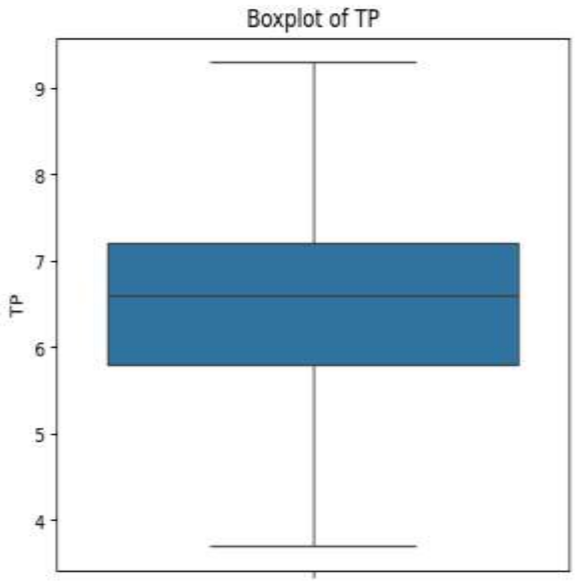
Aspartate Aminotransferase (AST): AST levels range from 10 to 180, with an average of about 65.2 and a standard deviation of 54.1. Similar to ALT, elevated AST levels can indicate liver inflammation or damage.



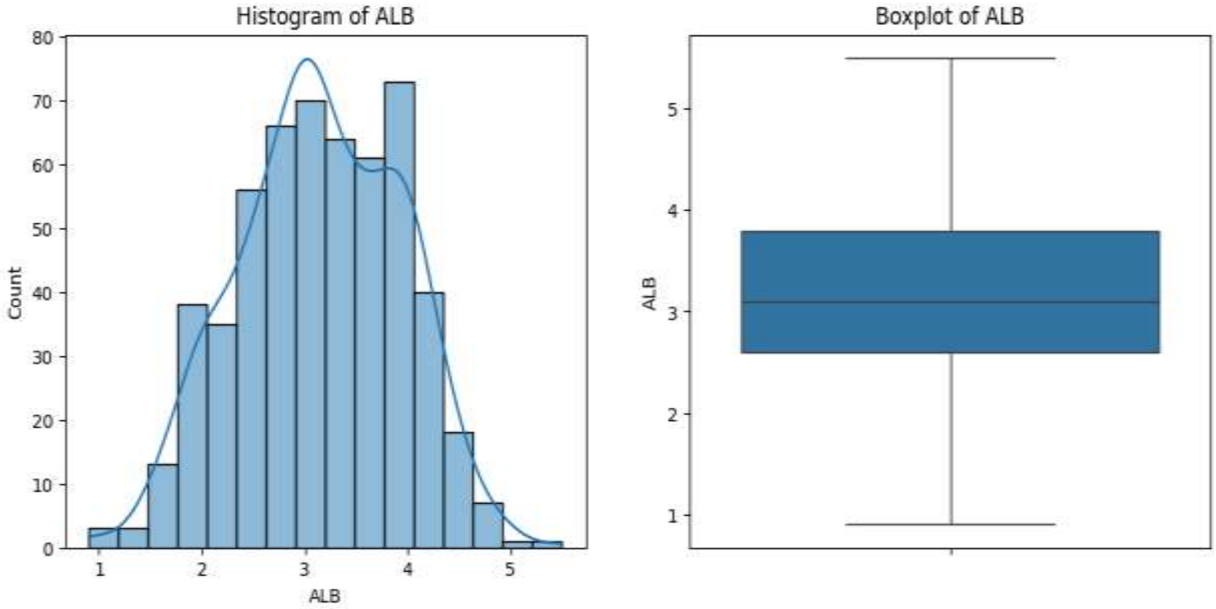


Total Proteins (TP): TP values range from 3.7 to 9.3, averaging around 6.48 and with a standard deviation of 1.06. Total proteins include albumin and globulin, essential for various bodily functions. Variability in TP levels can suggest different nutritional or health statuses among the patients.

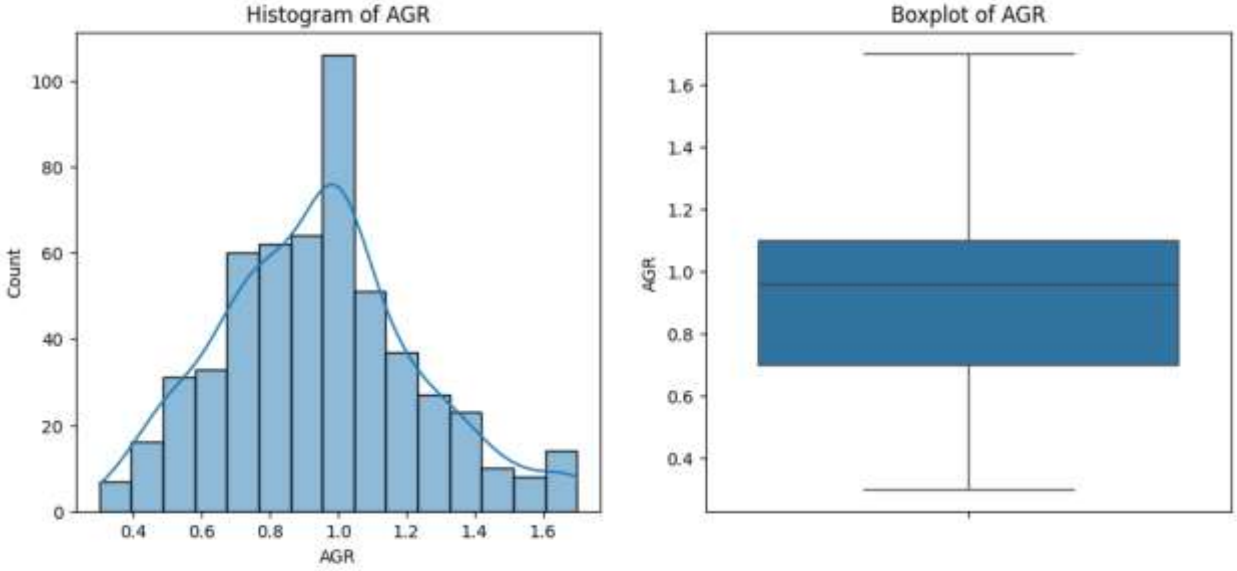




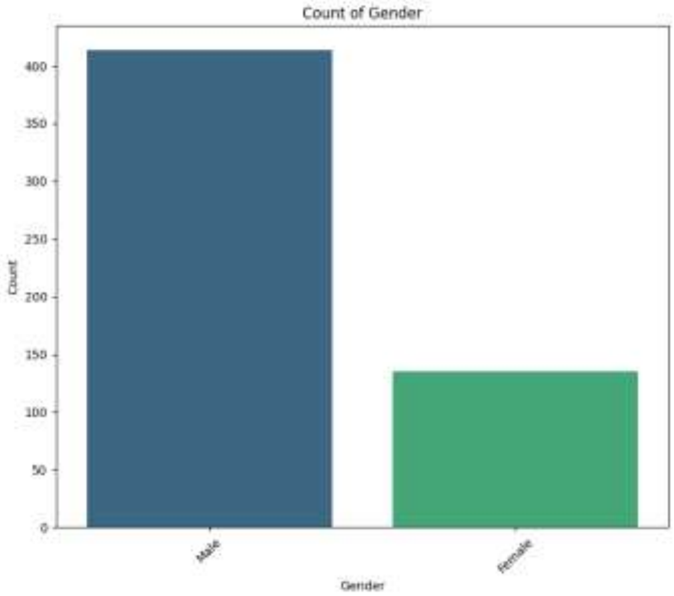
Albumin (ALB): Albumin levels range from 0.9 to 5.5, with a mean of 3.14 and a standard deviation of 0.79. The liver makes albumin, a crucial protein, and low levels can indicate liver disease or malnutrition.

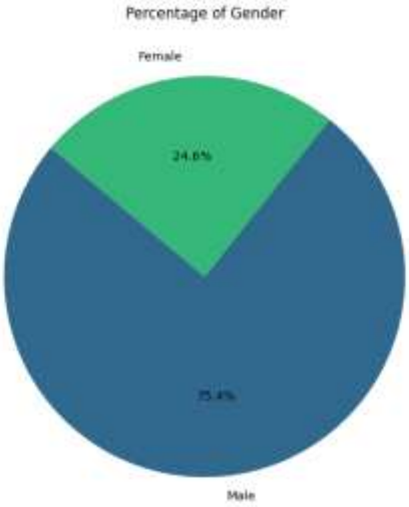


Albumin and Globulin Ratio (AGR): AGR values range from 0.3 to 1.7, with an average of 0.95 and a standard deviation 0.30. This ratio assesses liver function and immune status, with deviations from the norm indicating potential health issues.

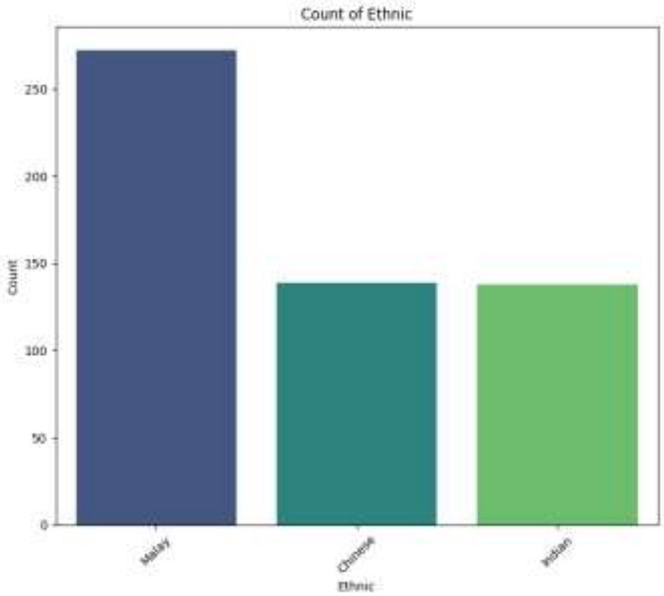


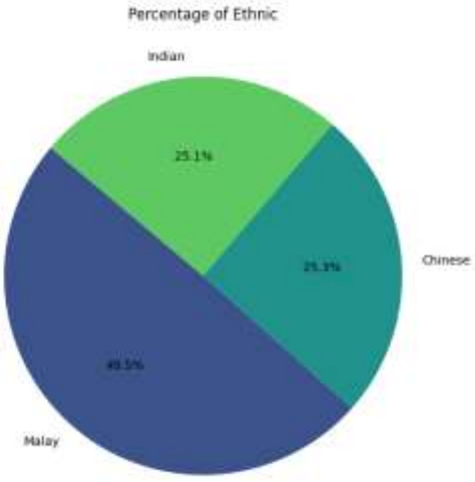
Gender: The dataset comprises a higher proportion of male patients compared to female patients. Specifically, there are 417 male patients, accounting for 75.4% of the total, and 136 female patients, making up the remaining 24.6%. This is visually represented in the bar plot and pie chart below.



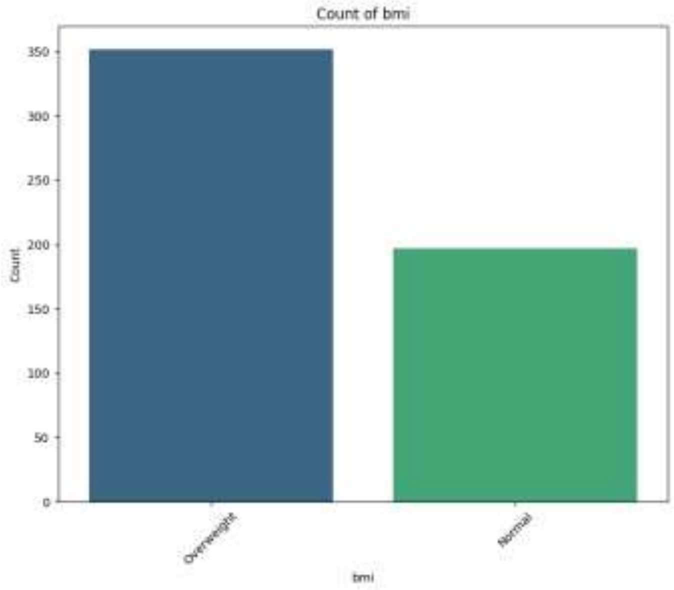


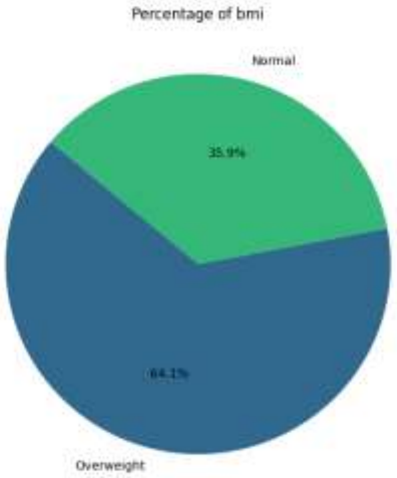
Ethnicity: The distribution of ethnicities in the dataset shows that Malay individuals form the largest group, followed by Chinese and Indian individuals. Specifically, there are 274 Malay patients (49.5%), 140 Chinese patients (25.3%), and 139 Indian patients (25.1%). This distribution is depicted in the following bar plot and pie chart.



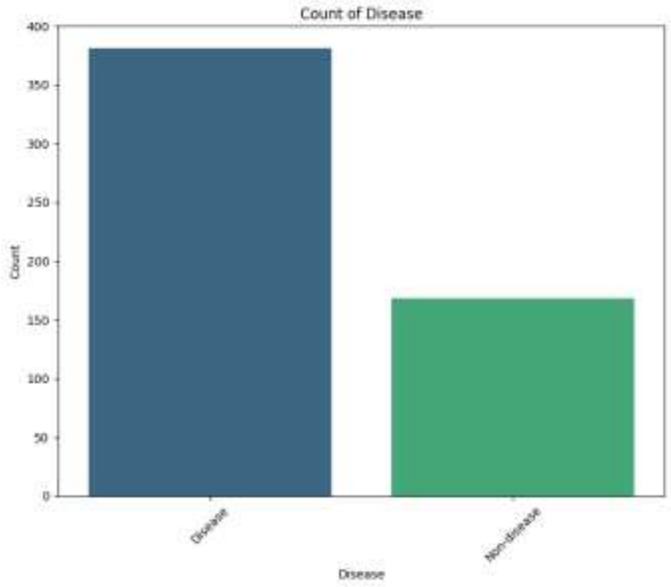


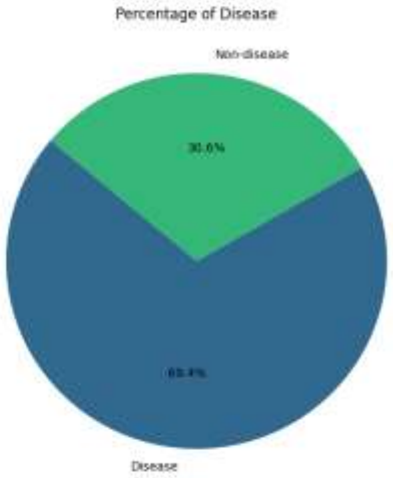
Body Mass Index (BMI): BMI values are categorized as 1 for Overweight and 2 for Normal. Approximately 64% of patients have a normal BMI, while 36% are overweight. This categorical variable provides insight into the patients' weight distribution, which can be relevant for liver disease risk factors.





Disease Status: The target variable indicates whether a patient has liver disease (1 for disease, 2 for no disease). In this dataset, about 69.4% of patients have liver disease, while 30.6% do not have liver disease. This distribution helps understand the class imbalance, which is crucial for model training and evaluation.





2.0 METHODOLOGY

The methodology section outlines the systematic steps to preprocess the data, train the machine learning models, and evaluate their performance. This ensures that the analysis is reproducible, transparent, and based on robust statistical techniques. The methodology is divided into several key stages: data preprocessing, algorithm design and implementation, and model evaluation.

2.1 Data Preprocessing

1. Loading and Initial Inspection:

The data preprocessing commences with data loading and initial inspection. This phase entails importing the data and examining the initial rows to comprehend its structure. Subsequently, the data types of each column are meticulously assessed to identify any necessary conversions for further analysis.

2. Removing Unnecessary Columns:

After the initial inspection, the Patient_ID column is deliberately removed from the dataset. While this column is vital in patient identification, it holds no predictive value in the context of analysis. This removal streamlines the data and eliminates irrelevant information.

3. Handling Missing Values:

Missing values, particularly within the AGR column, necessitate further attention during preprocessing. To ensure data integrity and facilitate subsequent analyses, rows containing missing values are strategically removed. This approach results in a refined dataset with a potentially reduced number of rows but guarantees complete information within the remaining data points.

4. Encoding Categorical Variables:

The next stage of data preprocessing focuses on encoding categorical variables. One such variable, gender, is transformed to facilitate seamless integration with machine learning algorithms. This process involves assigning numerical representations to distinct categories. For instance, the category "male" is encoded as 0, while "female" is encoded as 1. This conversion enhances the suitability of the data for machine learning applications.

5. Identifying and Handling Outliers:

Outliers in numerical columns are identified using the Interquartile Range (IQR) method. This involves calculating the first and third quartiles (Q1 and Q3) and determining the IQR (Q3 - Q1). Values outside 1.5 times the IQR above Q3 or below Q1 are considered outliers. Outliers are capped to the nearest acceptable values within the range defined by the IQR method. This prevents extreme values from skewing the analysis and model training.

6. Calculating Descriptive Statistics:

Descriptive statistics such as mean, standard deviation, minimum, maximum, and quartiles are calculated for each numerical variable. These statistics comprehensively summarize the data distribution, central tendencies, and variability.

7. Data Visualization:

Bar plots and pie charts are created to visualize the distribution of categorical variables like gender, ethnicity, BMI, and disease status. Histograms and boxplots are generated for numerical variables to illustrate their distribution and highlight potential outliers.

8. Dataset Splitting

After inputting missing values, the dataset is split into two groups, stratified by the outcome variable: 80% for training and testing the model and 20% as a hold-out set for validating the model’s performance.

9. Feature Scaling:

Feature scaling is performed to standardize the numerical features with a mean of 0 and a standard deviation of 1. This ensures that all features contribute equally to the model training process and prevents any single feature from disproportionately influencing the model due to its scale.

10. Handling Class Imbalance:

The class distribution of the target variable (Disease) is checked. Given the imbalance (more patients without liver disease than with), the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training set. SMOTE generates synthetic samples for the minority class, balancing the dataset and improving the model's ability to learn from both classes.

2.2 Model-Building Strategy & Algorithms

This study adapts model-building strategies from previous research to ensure reliable and robust results in developing a predictive model for liver disease. This process aims to classify patients into 'liver disease present' or 'liver disease absent' groups based on their biochemical markers and clinical profiles.

There are 7 ML algorithms are used to predict liver disease:

1. Logistic Regression

Logistic Regression is a statistical method for analyzing a dataset in which one or more independent variables determine an outcome. The outcome is a binary variable (yes/no, 0/1). It models the probability that a given input point belongs to a particular class. In the context of liver disease prediction, logistic regression can estimate a patient's likelihood of liver disease based on various predictors such as age, gender, ethnicity, BMI, and blood test results.

2. Support Vector Machine (SVM)

Support Vector Machine is a supervised learning model that analyzes data for classification and regression analysis. SVM constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification. The goal is to find a hyperplane that best divides the data into classes. For predicting liver disease, SVM would attempt to create a decision boundary that separates patients with liver disease from those without based on the features provided.

3. K-Nearest Neighbors (k-NN)

K-Nearest Neighbors is a simple, non-parametric, and instance-based learning algorithm. The principle behind k-NN is to classify a data point based on its neighbors' classification. It looks at the 'k' closest training examples in the feature space and assigns the class most common among its k nearest neighbors. For liver disease prediction, this would involve comparing new patient data to those of patients with known liver disease status and predicting the new patient's status based on the most common outcome among the closest matches.

4. Decision Tree

A Decision Tree is a non-parametric supervised learning method for classification and regression. The model splits the data into subsets based on the value of input features. Each node in the tree represents a feature (attribute), each branch represents a decision rule, and each leaf represents an outcome. For liver disease prediction, a decision tree would create a series of binary splits on different features (like blood test results) to classify whether a patient has liver disease.

5. Random Forest

Random Forest is an ensemble learning method that constructs many decision trees during training and outputs the mode of the individual trees' classes (classification) or mean prediction (regression). It improves the predictive accuracy and controls over-fitting. For liver disease prediction, a Random Forest model would aggregate the predictions of many decision trees to improve the robustness and accuracy of the prediction.

6. Extreme Gradient Boosting (XGBoost)

XGBoost is an implementation of gradient-boosted decision trees designed for speed and performance. It builds the model in a stage-wise fashion, and it generalizes the model by optimizing a differentiable loss function. For liver disease prediction, XGBoost can handle various types of data and can be tuned to provide high predictive performance by learning from the errors of previous models in the boosting sequence.

7. Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient, with faster training speed, higher efficiency, lower memory usage, and better accuracy. LightGBM uses a leaf-wise rather than a level-wise approach, making it more efficient. LightGBM can handle large datasets and numerous features effectively for predicting liver disease, making it a strong candidate for such tasks.

2.3 Model Evaluation

1. Sensitivity Analysis

To address potential biases introduced by SMOTE and imputation, a sensitivity analysis is conducted by retraining models without SMOTE and, for specific models like XGBoost, without imputing missing values. Sensitivity analysis is performed by adjusting the parameters of each model to determine their robustness and optimal performance settings.

2. Feature importance

Feature importance is analyzed for models that provide this capability, such as Random Forest and Logistic Regression. This analysis helps to identify which features contribute most significantly to the prediction of liver disease, offering insights into the underlying factors that influence the models' decisions.

3. Cross-Validation

Cross-validation is performed to assess each model's stability and generalizability. This involves partitioning the training set into several folds and training the model on different subsets while validating the remaining data.

4. Performance Metrics Across Different Thresholds

For models like Logistic Regression and Random Forest, performance metrics (sensitivity, specificity, and accuracy) are calculated across different decision thresholds. This helps in understanding how changes in the threshold affect the model's performance.

5. Model Training and Hyperparameter Tuning

Each model is trained using stratified k-fold cross-validation. Hyperparameter tuning is performed using grid search to find the optimal settings. The ROC-AUC score is used to identify the best hyperparameters, considering the trade-off between true positive and false positive rates.

6. Model Performance Evaluation

The best model for each algorithm is evaluated on the hold-out dataset using metrics such as mean ROC-AUC Score, accuracy, precision, recall, sensitivity, specificity, and F1 score. A confusion matrix is also presented to illustrate the models' performance.

2.4 Results Interpretation

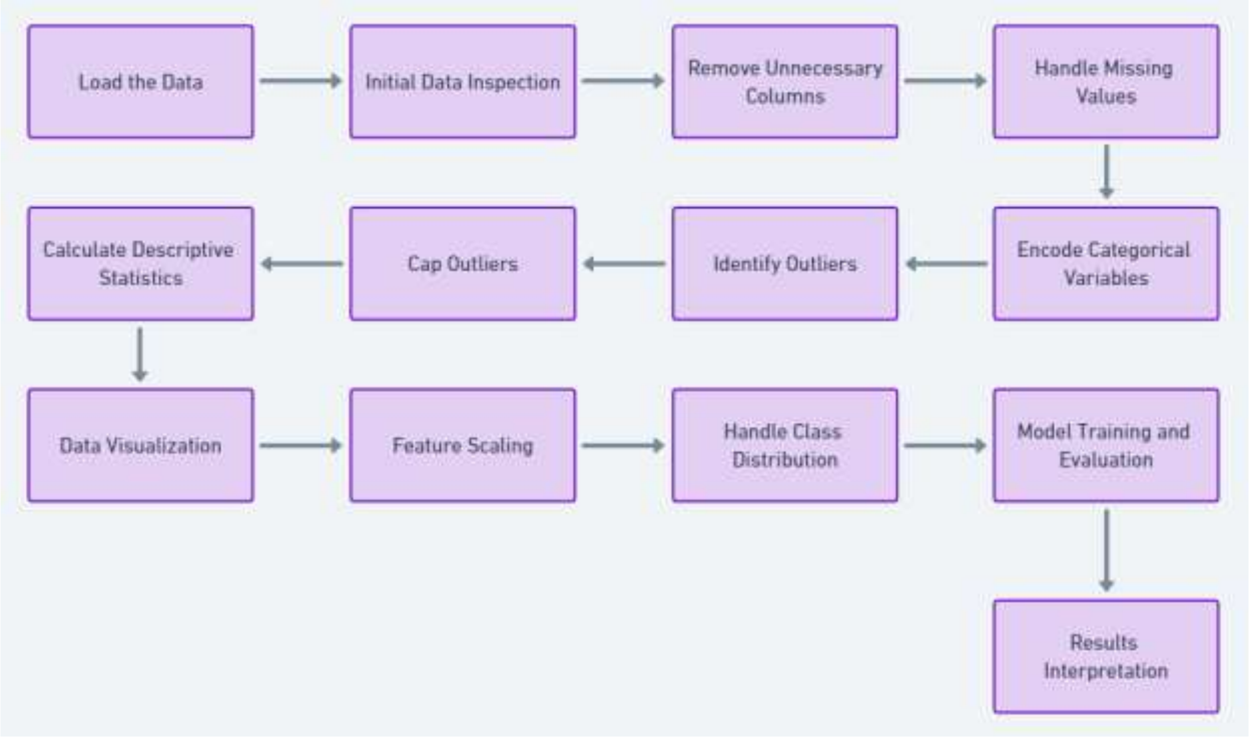
In this study, the performance of each model in predicting liver disease was rigorously interpreted, focusing on accuracy and other critical metrics. Comparisons were made to identify the most effective model, utilizing performance metrics such as Mean ROC-AUC Score, Accuracy, Precision, Recall, F1-Score, RMSE, and MARD. These metrics provided a comprehensive understanding of each model's strengths and weaknesses in predicting liver disease.

The comparative analysis extended beyond internal comparisons, incorporating a review of existing literature to validate our findings. This step highlighted the strengths and weaknesses of each approach in the context of established research, ensuring a thorough evaluation of each model's performance.

The final step involved identifying the best-performing model by assessing its performance across all metrics and its robustness in sensitivity analysis and cross-validation. The model that demonstrated consistent high performance in these evaluations was identified as the most effective for predicting liver disease in our dataset.

This comprehensive methodology, encompassing data preparation, rigorous model evaluation, and thorough result validation, ensures that the findings are robust and reliable. As a result, this research provides valuable insights into the prediction of liver disease, contributing to the field's understanding and potentially guiding future research and clinical applications.

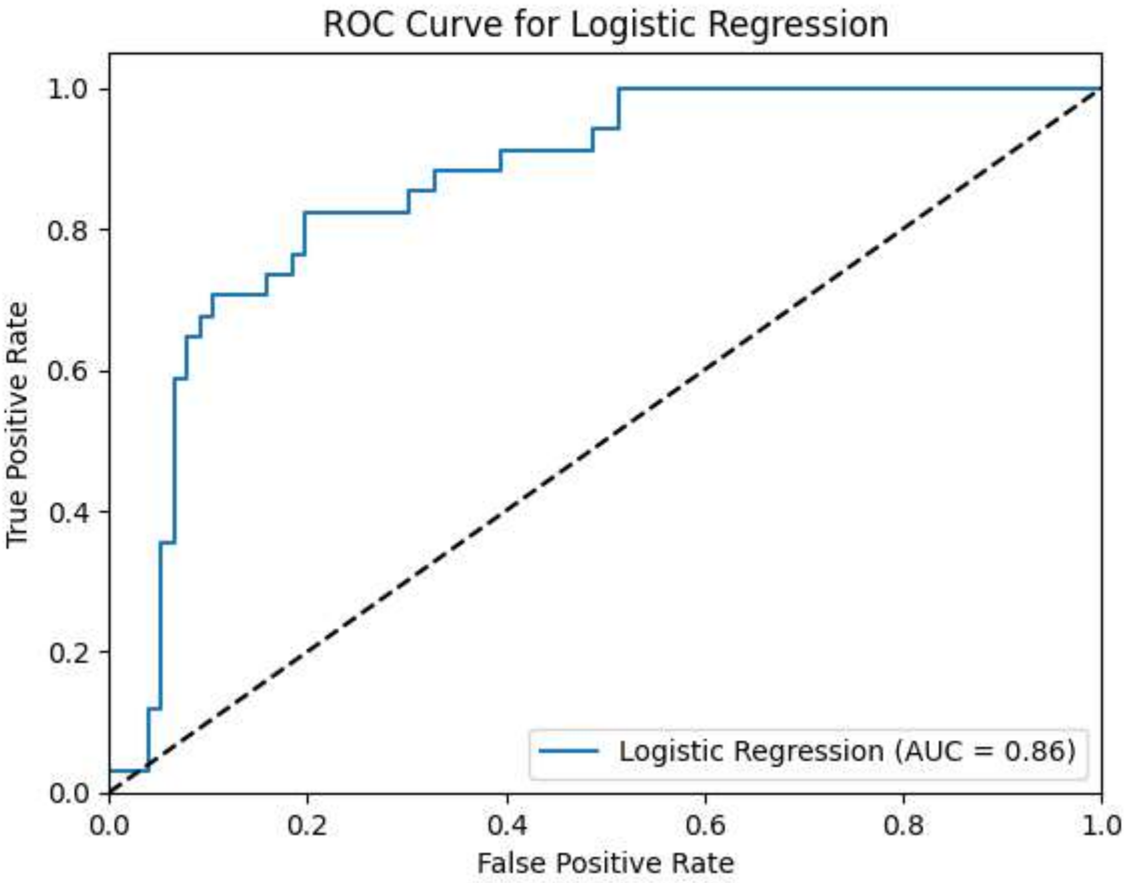
Figure: Flow Diagram of Data Preprocessing, Analysis, and Model Evaluation

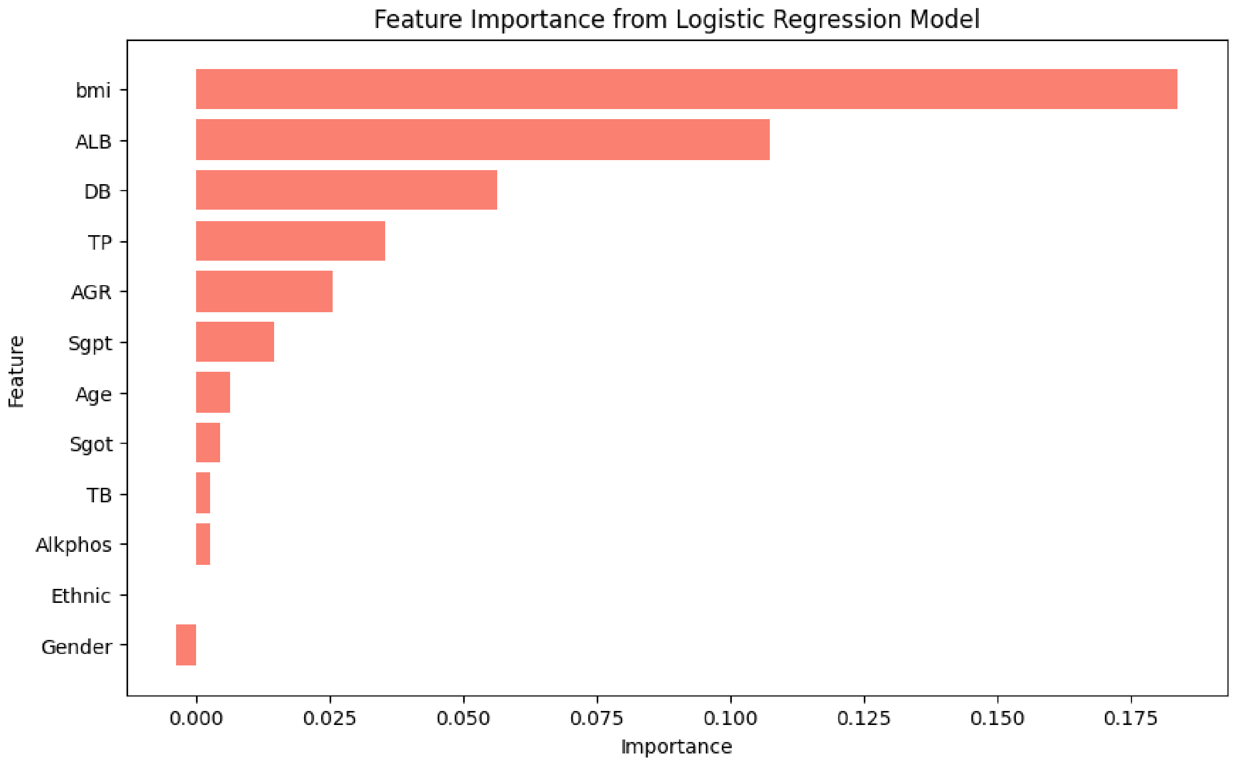


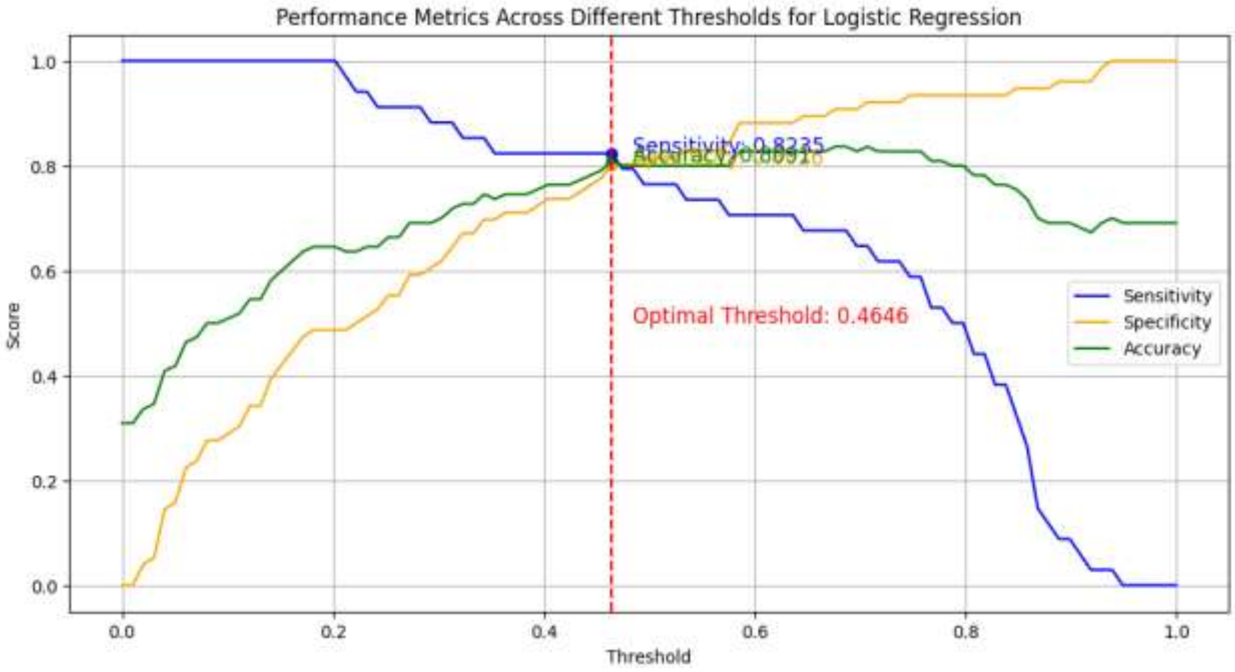
3.0 RESULTS

The analysis involved training and evaluating seven machine learning models: Logistic Regression, Random Forest, lightGBM, K-nearest neighbors (KNN), XGBoost, Support Vector Machine (SVM), and Decision Tree. Each model's performance was assessed using various metrics, including confusion matrix, precision, recall, F1-score, ROC-AUC, RMSE, and MARD.

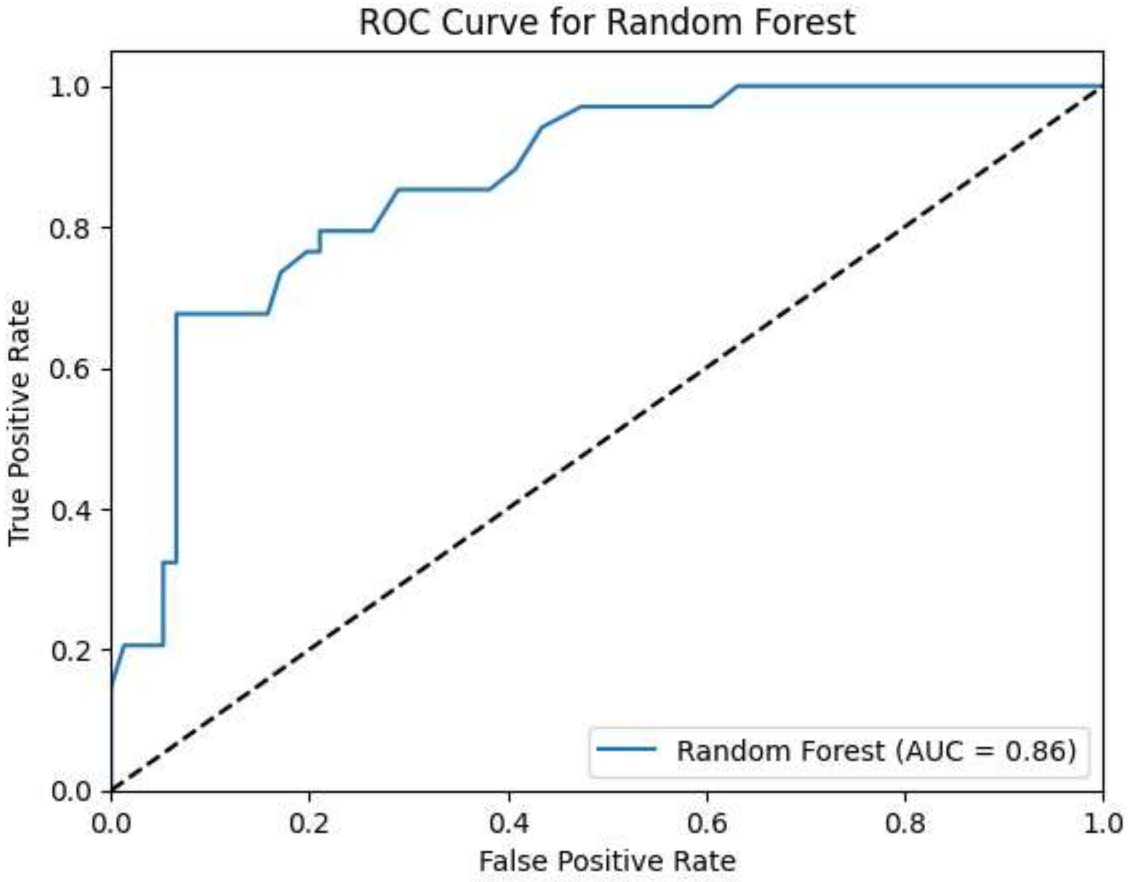
Logistic Regression achieved an accuracy of 0.80, with a precision of 0.89 for patients with liver disease (Class 0) and 0.65 for those without liver disease (Class 1). The recall was 0.82 for Class 0 and 0.76 for Class 1, resulting in F1 scores of 0.85 and 0.70, respectively. The ROC-AUC score was 0.86. Sensitivity analysis, which varied the regularization strength (C) from 0.1 to 10, showed that the ROC-AUC scores ranged from 0.86 to 0.87, with the optimal performance at C=1.0. Logistic Regression maintains a good balance between all performance metrics, with a notable ROC-AUC score and accuracy. It is effective for applications where interpretability and balance are essential. Important features include BMI, ALB, and DB.

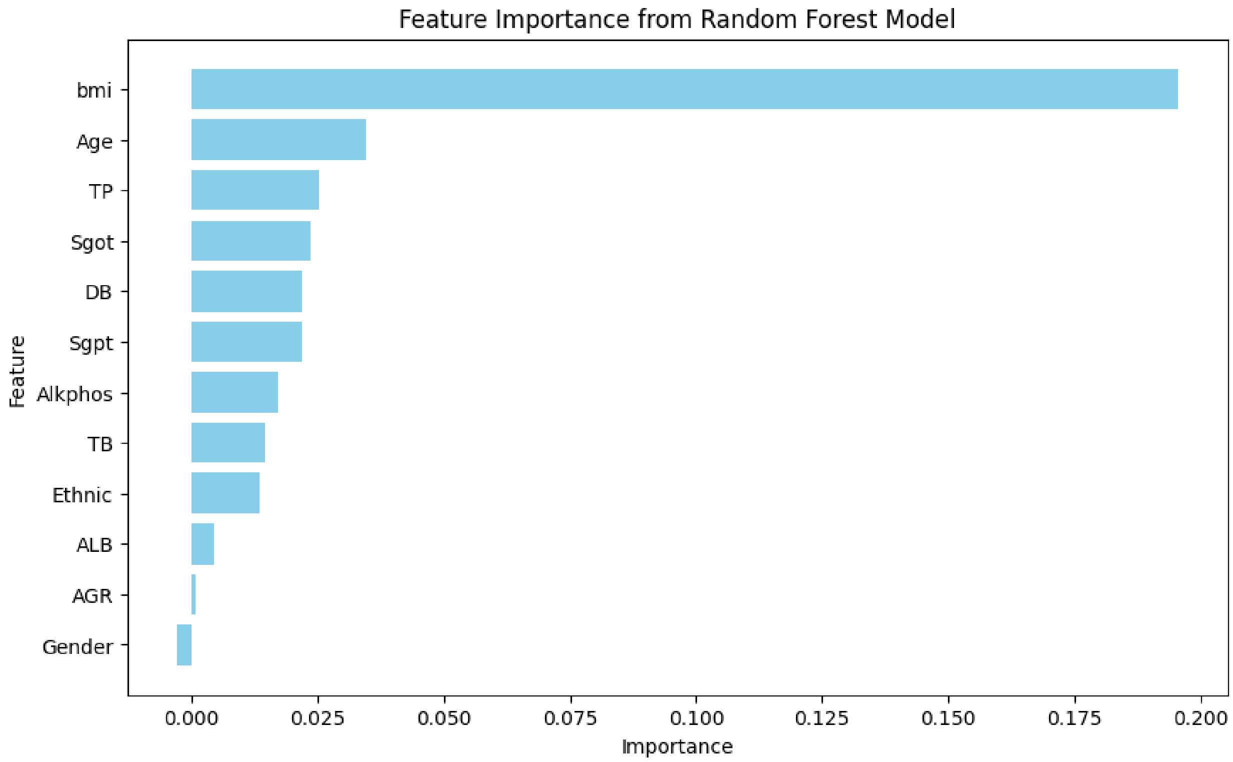


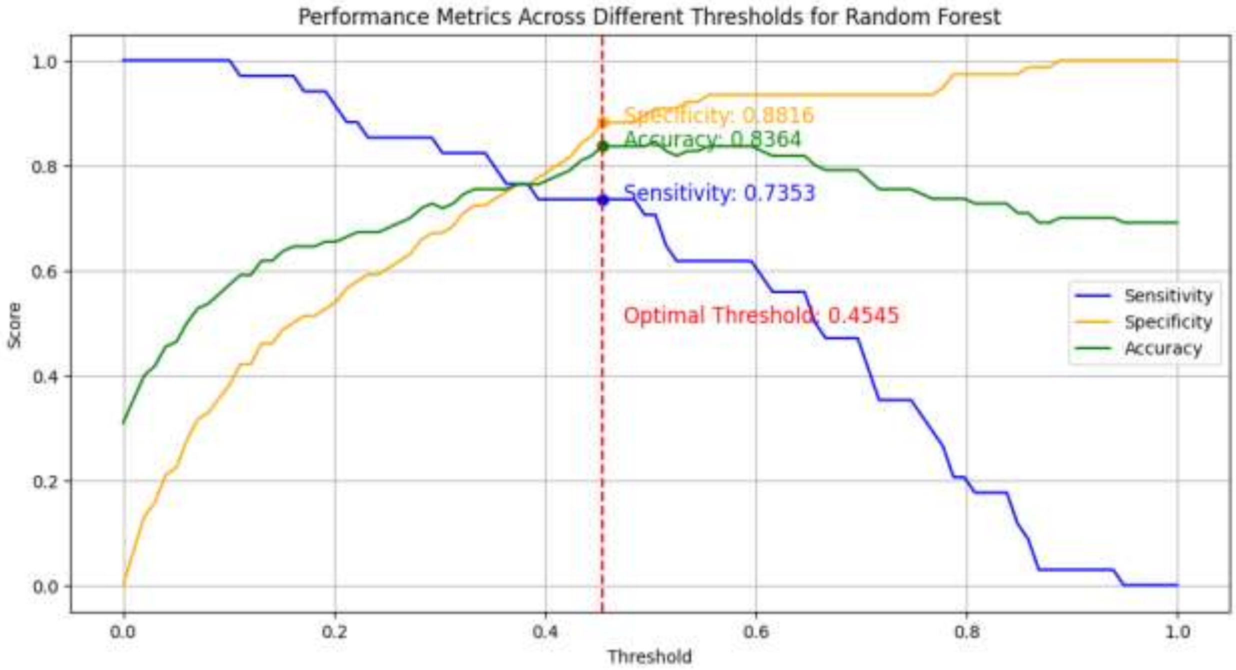

Feature Importance from Logistic Regression Model
Feature
bmi
ALB
DB
TP
AGR
Sgpt
Age
Sgot
TB
Alkphos
Ethnic
Gender
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance


Performance Metrics Across Different Thresholds for Logistic Regression
Score
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity: 0.8235
Accuracy: 0.8091
Optimal Threshold: 0.4646
Sensitivity
Specificity
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Threshold

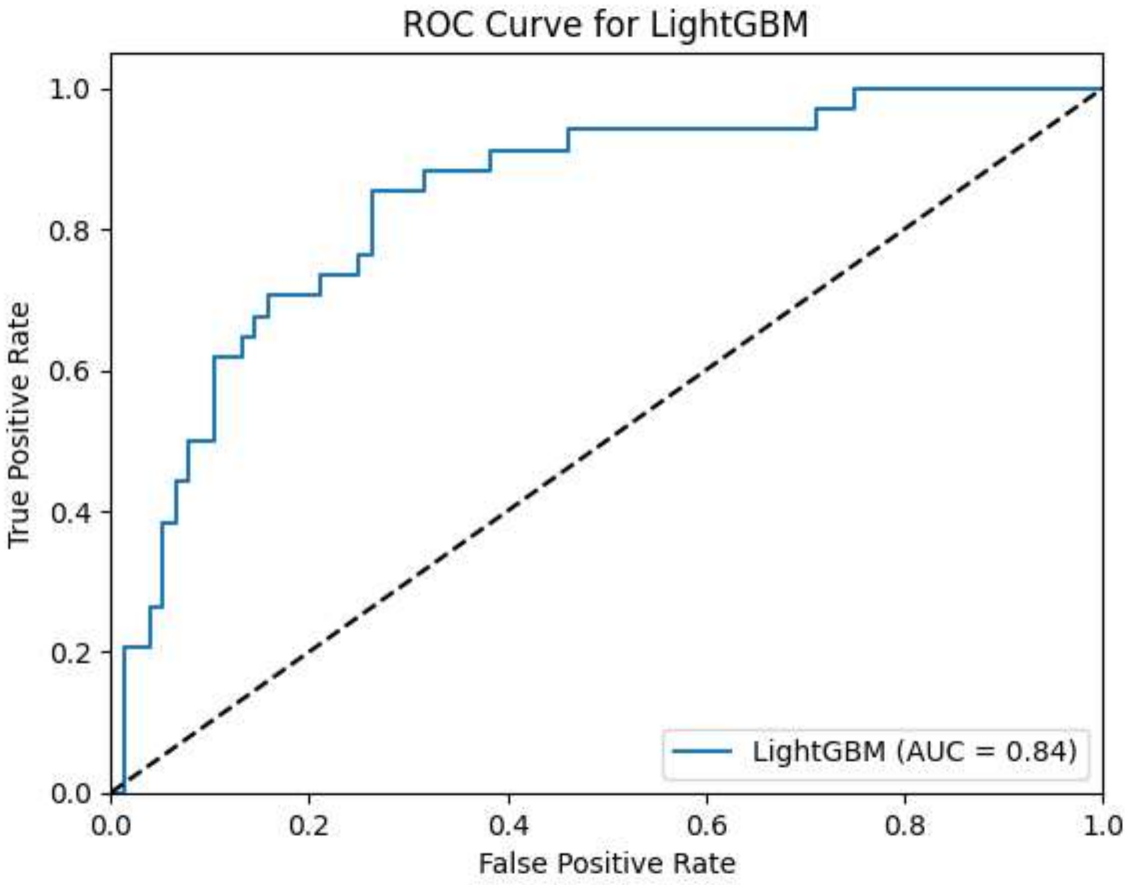
Random Forest emerged as the best-performing model with an accuracy of 0.84. It had a precision of 0.86 for Class 0 and 0.81 for Class 1, with a recall of 0.93 and 0.65, respectively. The F1 scores were 0.89 for Class 0 and 0.71 for Class 1. The model's ROC-AUC score was 0.86. Sensitivity analysis, varying the number of estimators from 100 to 300, showed ROC-AUC scores ranging from 0.85 to 0.86, with optimal performance at 100 estimators. The key features identified were BMI, SGOT, and Age. The Random Forest model performs excellently, with a high ROC-AUC score indicating discriminative solid ability. It has high accuracy and recall for Class 0 but lower recall for Class 1.

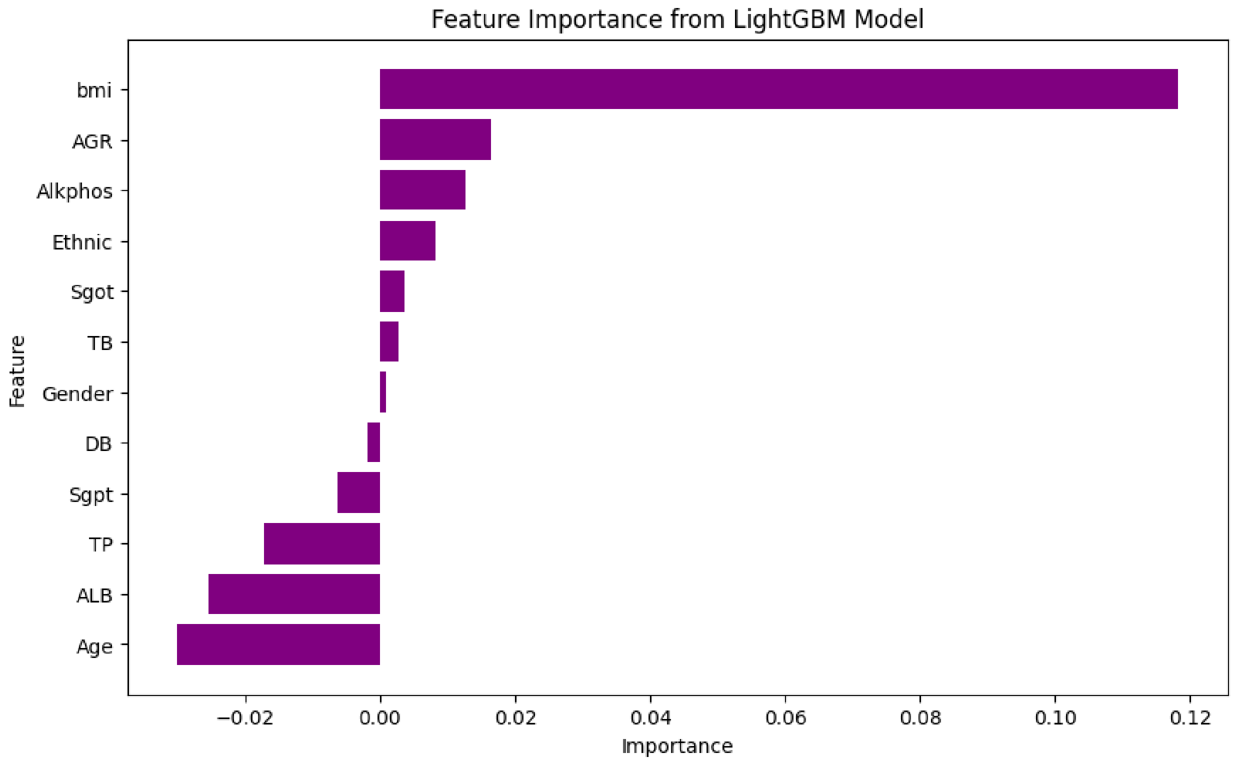


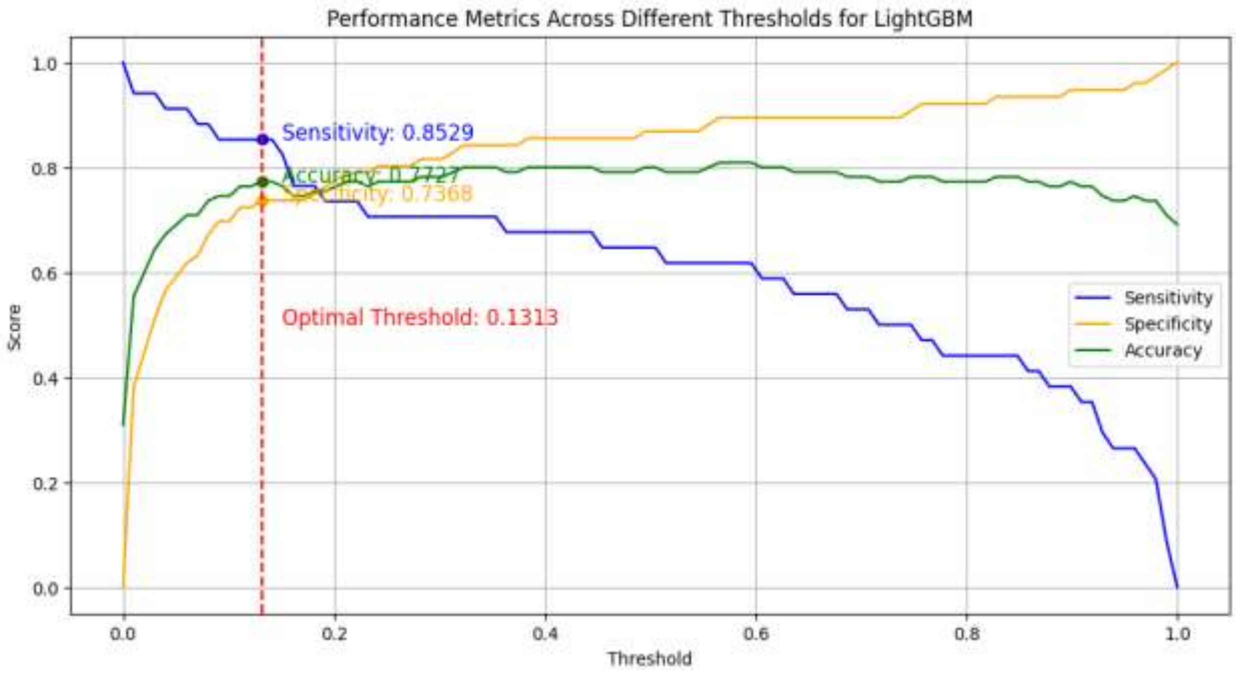

Feature Importance from Random Forest Model
Feature
bmi
Age
TP
Sgot
DB
Sgpt
Alkphos
TB
Ethnic
ALB
AGR
Gender
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance


Performance Metrics Across Different Thresholds for Random Forest
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Threshold
Specificity: 0.8816
Accuracy: 0.8364
Sensitivity: 0.7353
Optimal Threshold: 0.4545
Sensitivity
Specificity
Accuracy

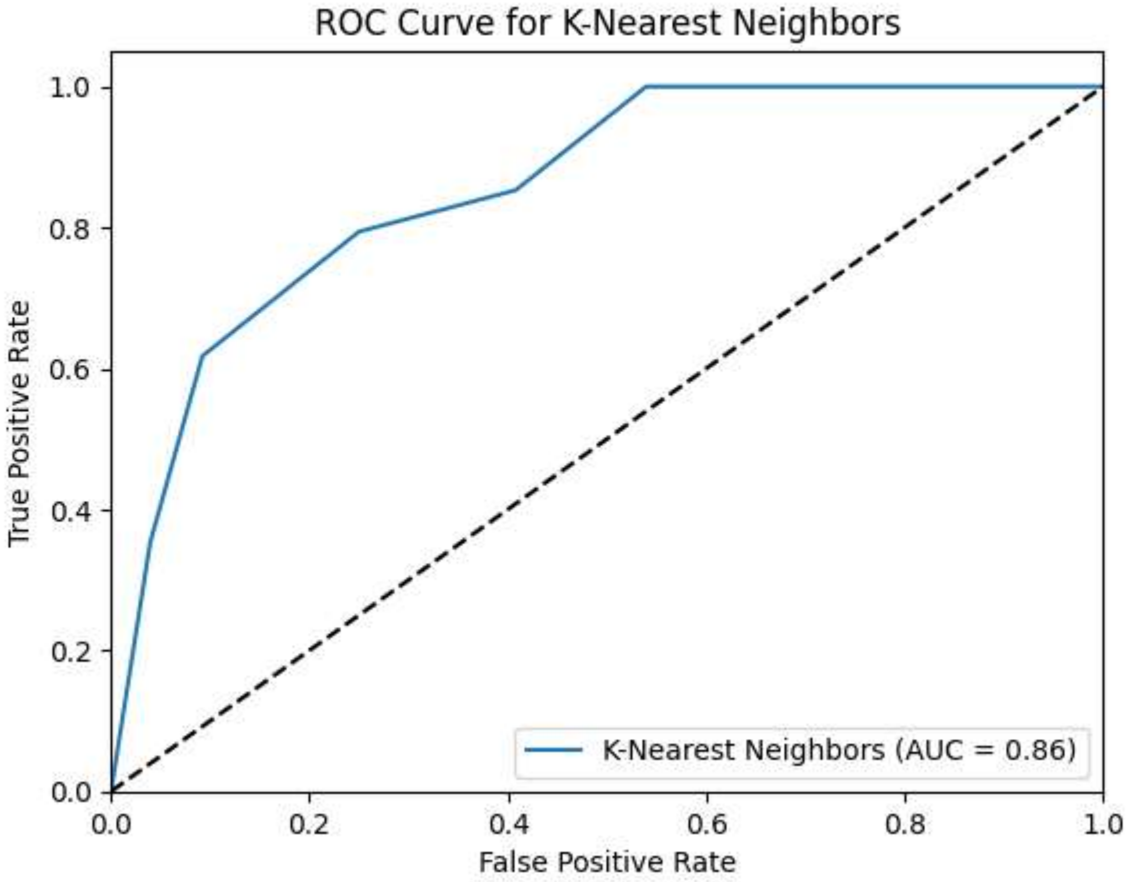
LightGBM showed strong results with an accuracy of 0.80 and a ROC-AUC score of 0.84. The precision for Class 0 was 0.85 and 0.69 for Class 1, with recall values of 0.87 and 0.65, respectively. The F1 scores were 0.86 for Class 0 and 0.67 for Class 1. Sensitivity analysis, varying the number of leaves and learning rates, identified the optimal parameters as 31 leaves and a learning rate of 0.1, achieving an ROC-AUC score of 0.84. LightGBM shows a high ROC-AUC score, good accuracy, and balanced precision and recall values. It performs well across most metrics, indicating it can be reliable in various scenarios. The most important features of this model are BMI, AGR, and ALKPHOS.

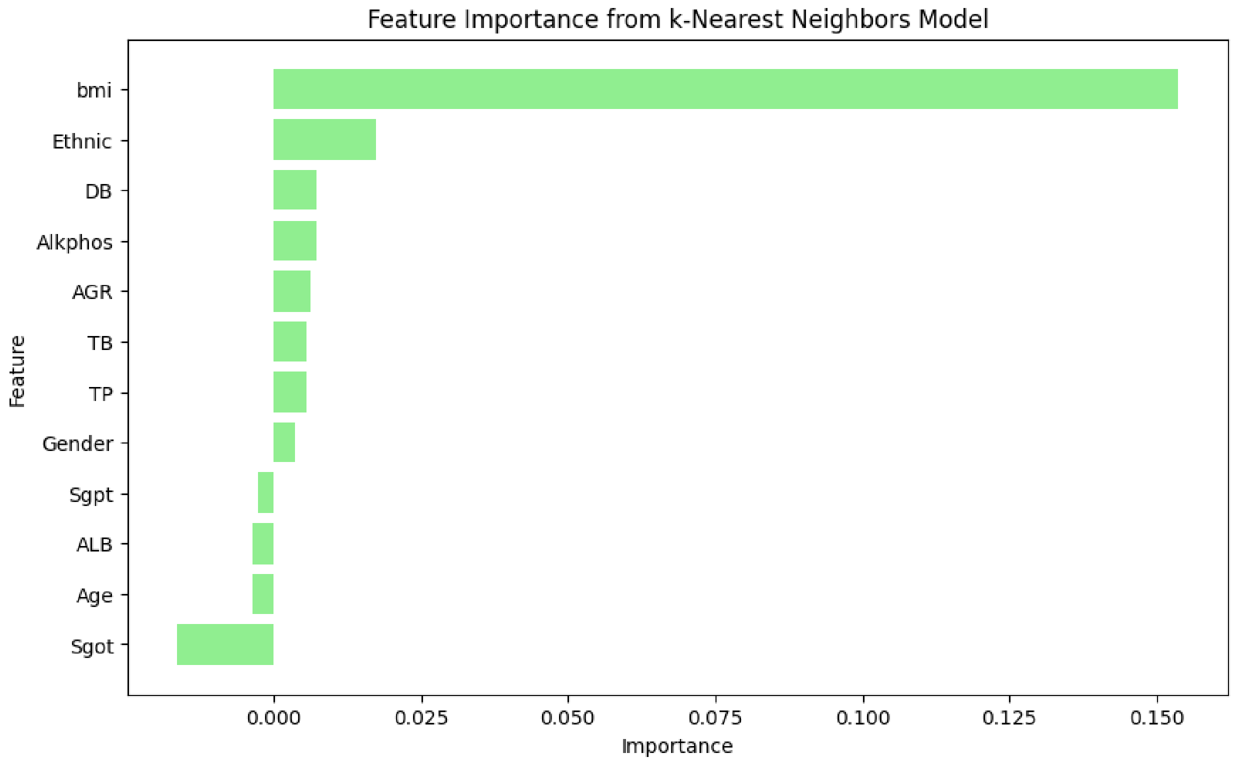


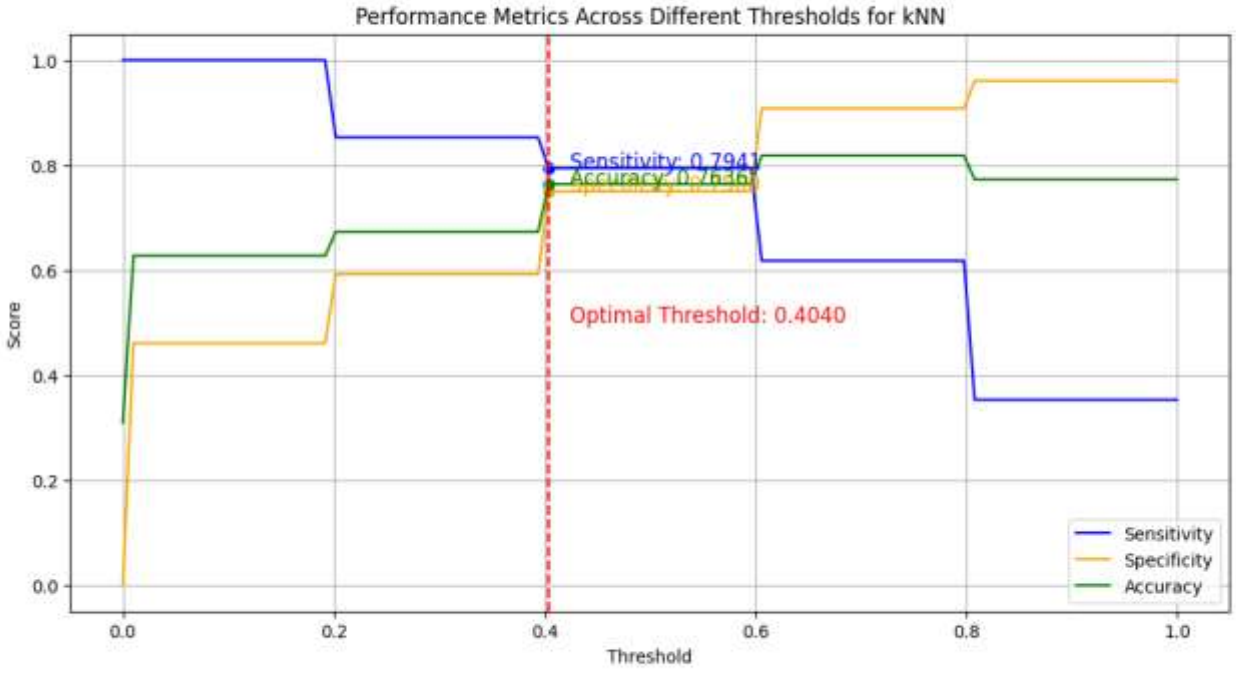

Feature Importance from LightGBM Model
Feature
bmi
AGR
Alkphos
Ethnic
Sgot
TB
Gender
DB
Sgpt
TP
ALB
Age
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance


Performance Metrics Across Different Thresholds for LightGBM
Score
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity: 0.8529
Accuracy: 0.7727
Specificity: 0.7368
Optimal Threshold: 0.1313
Sensitivity
Specificity
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Threshold

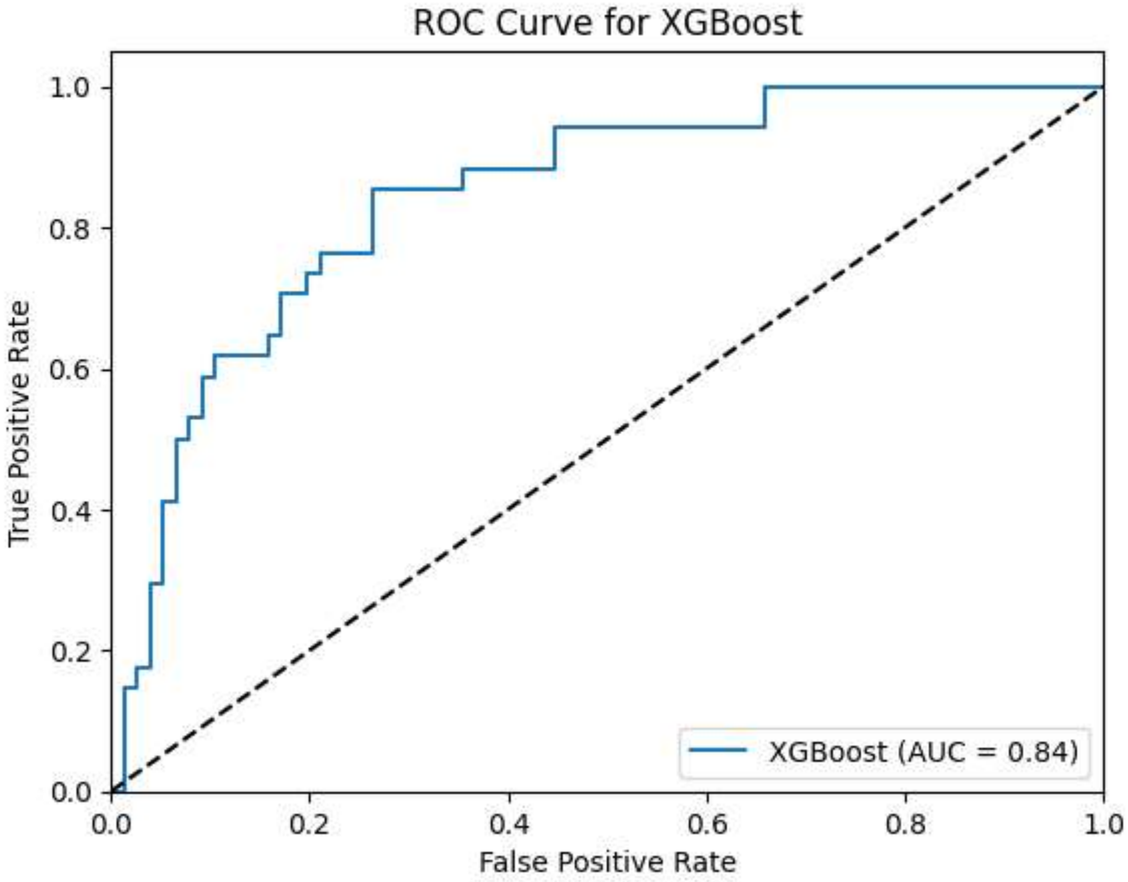
K-Nearest Neighbours (KNN) achieved an accuracy of 0.76, with precision values of 0.89 for Class 0 and 0.59 for Class 1. The recall was 0.75 for Class 0 and 0.79 for Class 1, resulting in F1 scores of 0.81 and 0.68, respectively. The ROC-AUC score was 0.83. Sensitivity analysis, testing different values of K, showed optimal performance at K=5, with a ROC-AUC score of 0.83. The KNN model has a strong ROC-AUC score and is particularly good at identifying the positive class, as evidenced by its recall for class 1. Key features for this model include BMI, ethnicity, and DB.

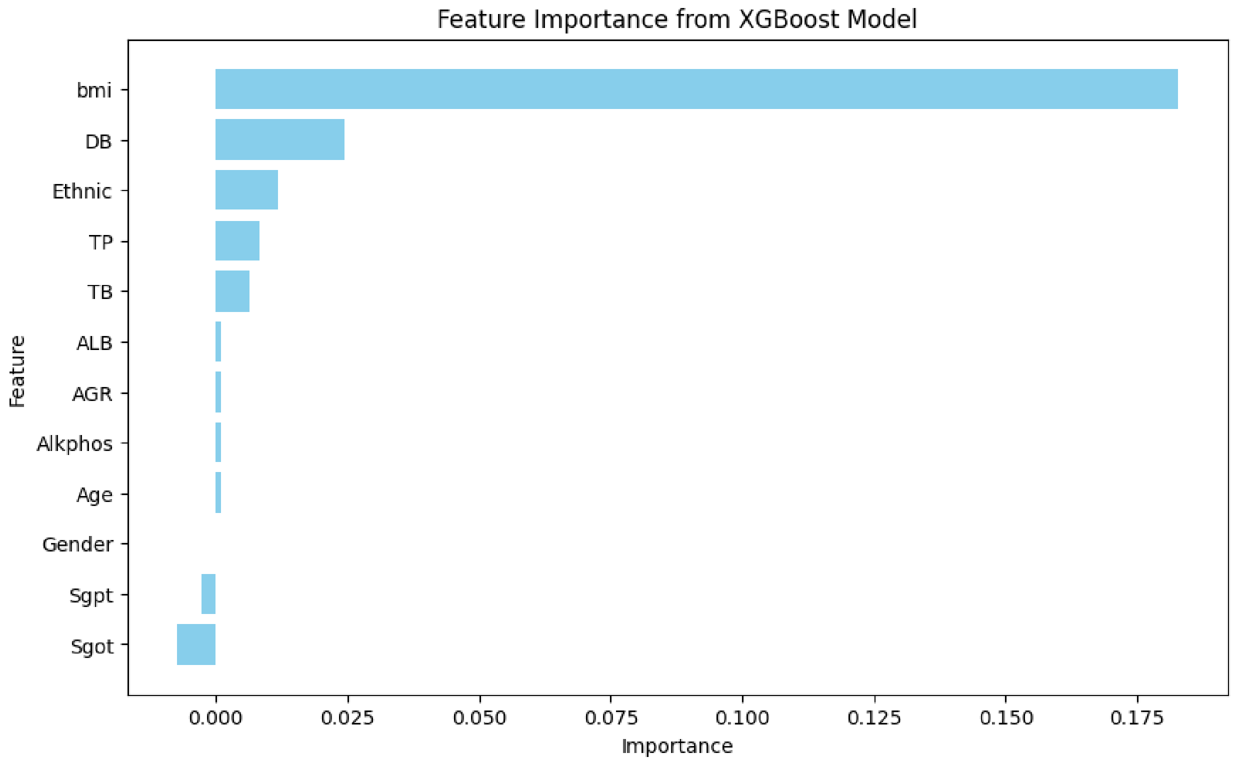


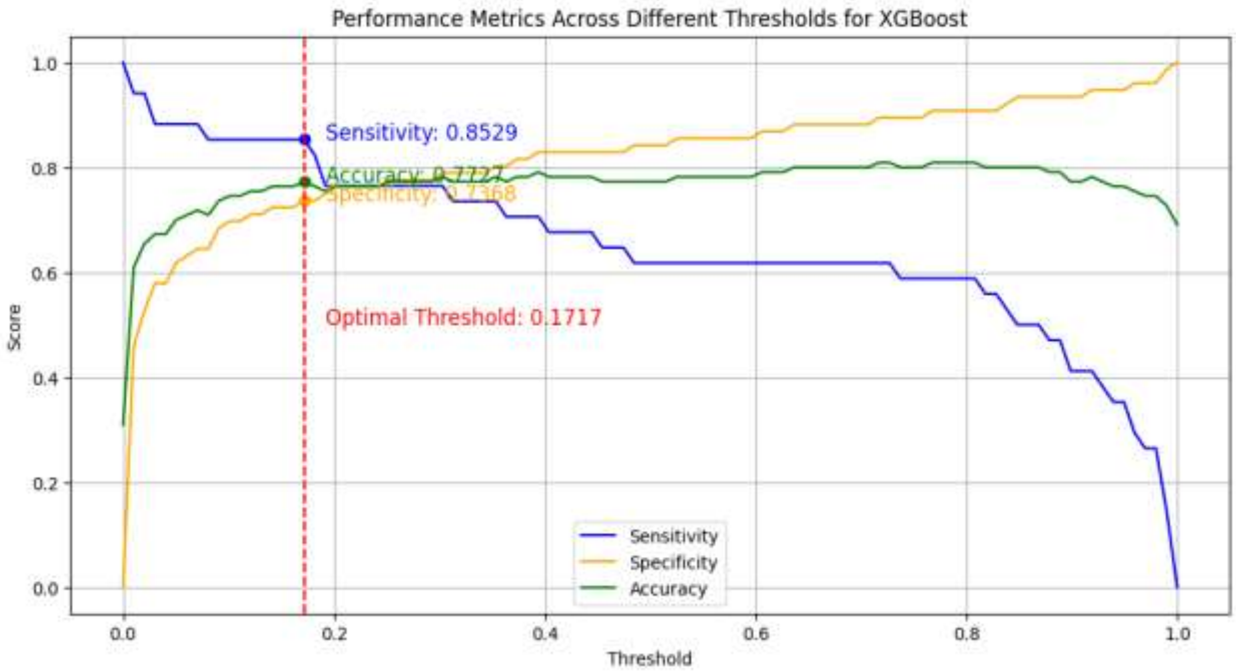

Feature Importance from k-Nearest Neighbors Model
Feature
bmi
Ethnic
DB
Alkphos
AGR
TB
TP
Gender
Sgpt
ALB
Age
Sgot
0.000
0.025
0.050
0.075
0.100
0.125
0.150
Importance


Performance Metrics Across Different Thresholds for kNN
Score
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity: 0.7941
Accuracy: 0.7636
Optimal Threshold: 0.4040
Sensitivity
Specificity
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Threshold

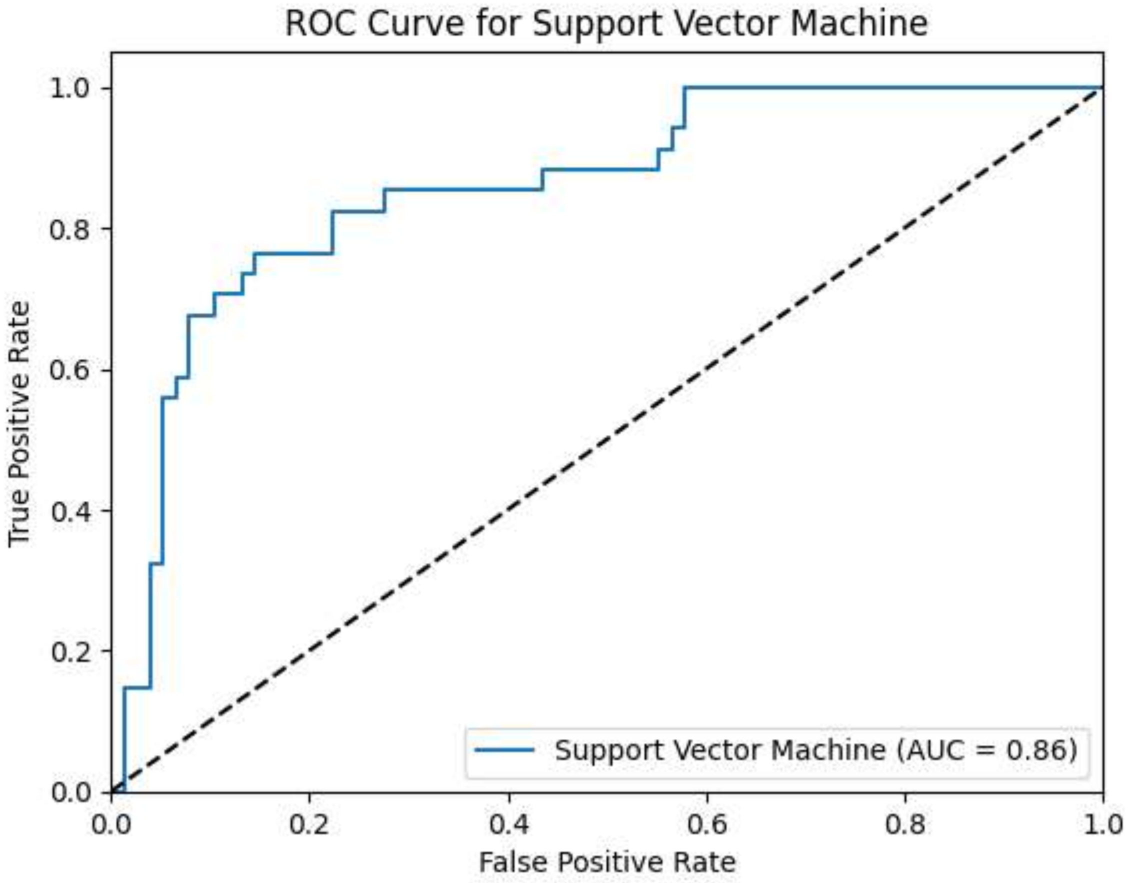
XGBoost achieved an accuracy of 0.77 and a ROC-AUC score of 0.83. The precision for Class 0 was 0.83 and 0.64 for Class 1, with recall values of 0.84 and 0.62, respectively. The F1 scores were 0.84 for Class 0 and 0.63 for Class 1. Sensitivity analysis, varying hyperparameters such as learning rate and the number of estimators, identified optimal performance with a learning rate of 0.1 and 100 estimators, resulting in a ROC-AUC score of 0.83. XGBoost has a solid ROC-AUC score and balanced performance metrics. It effectively handles both classes, making it a reliable choice for many applications. The top features are BMI, DB, and Ethnic.

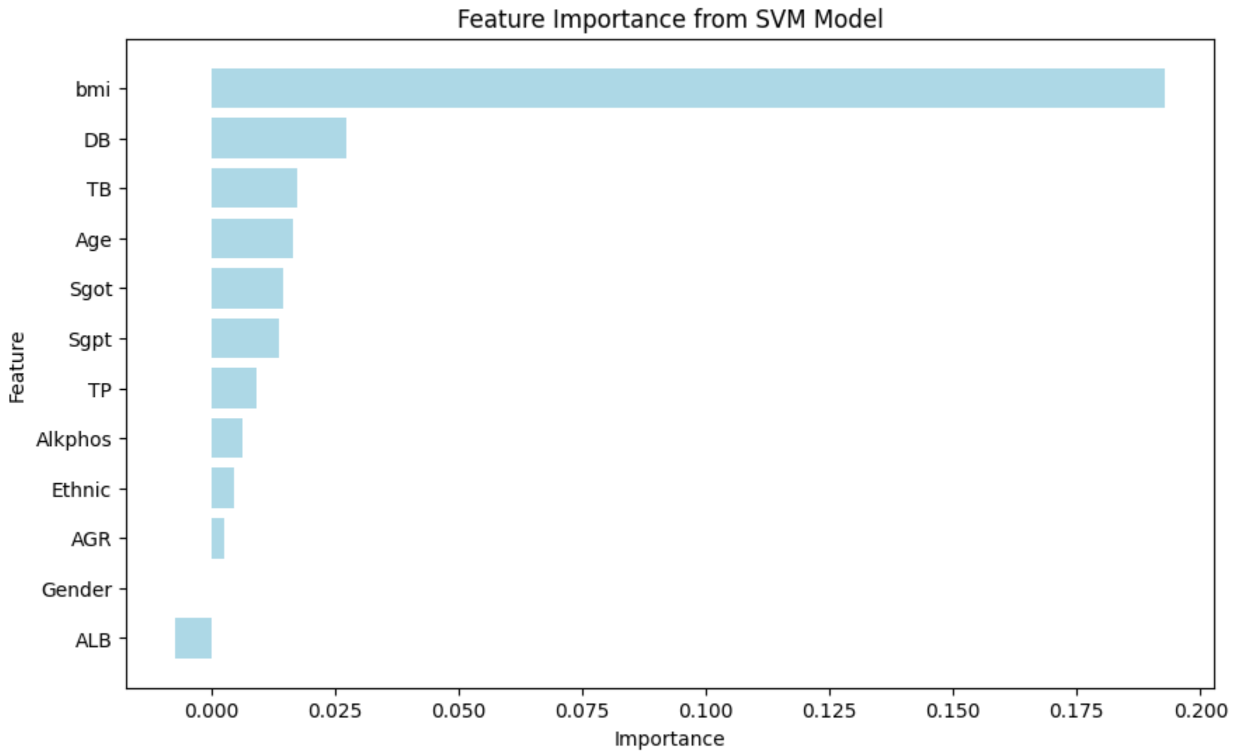


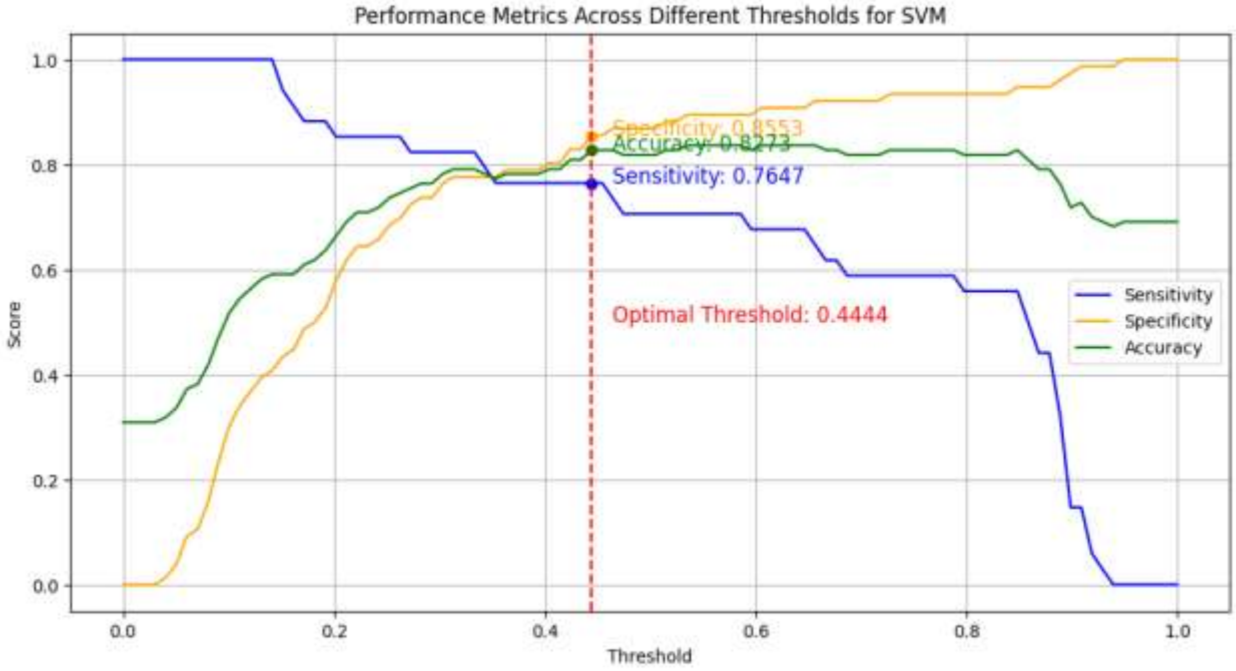

Feature Importance from XGBoost Model
Feature
bmi
DB
Ethnic
TP
TB
ALB
AGR
Alkphos
Age
Gender
Sgpt
Sgot
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance


Performance Metrics Across Different Thresholds for XGBoost
Score
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity: 0.8529
Accuracy: 0.7727
Specificity: 0.7368
Optimal Threshold: 0.1717
Sensitivity
Specificity
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Threshold

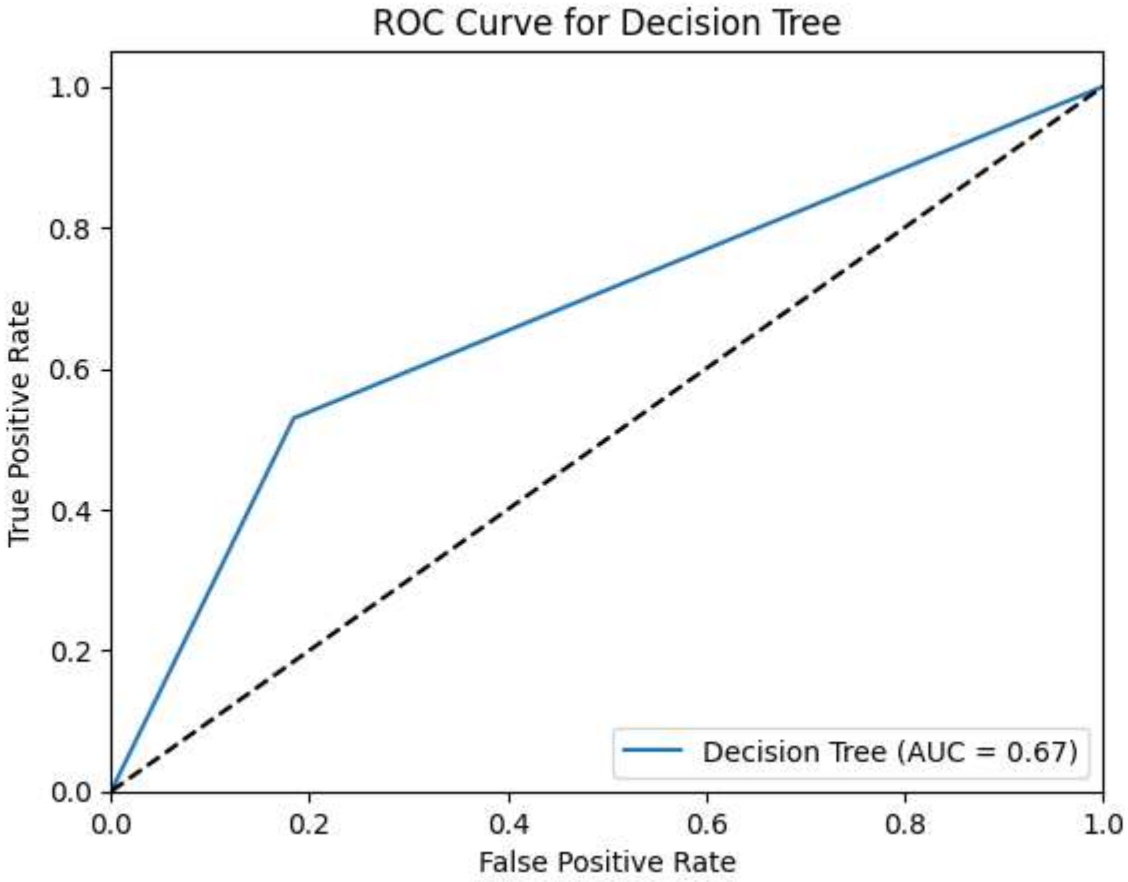
Support Vector Machine (SVM) demonstrated good accuracy at 0.83 and a high ROC-AUC score of 0.85. The precision was 0.87 for Class 0 and 0.73 for Class 1, with recall values of 0.88 and 0.71, respectively. The F1 scores were 0.88 for Class 0 and 0.72 for Class 1. Sensitivity analysis, testing different kernel functions and regularization parameters, showed optimal performance with a radial basis function (RBF) kernel and a regularization parameter of C=1, resulting in a ROC-AUC score of 0.85. The SVM model demonstrates high accuracy and well-balanced performance across all metrics. It is particularly effective for situations requiring high accuracy and balanced class identification. The main features influencing the model are BMI, DB, and TB.

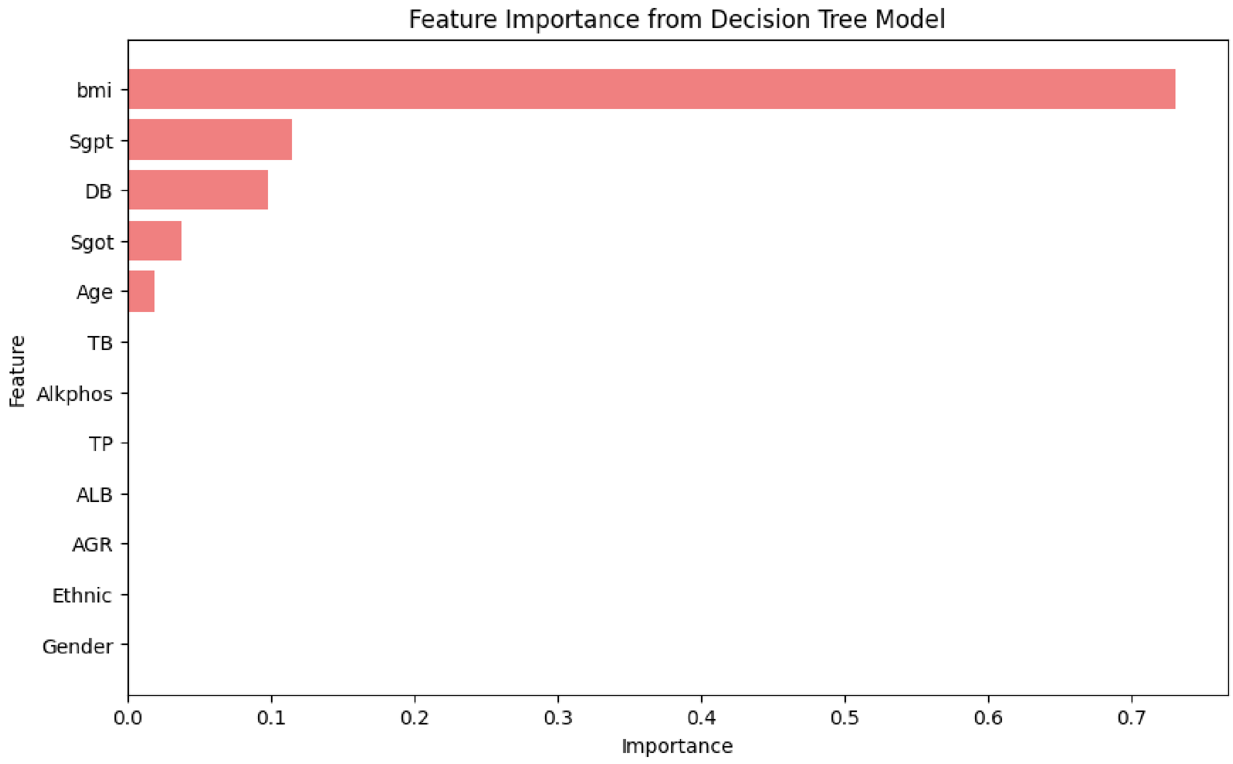


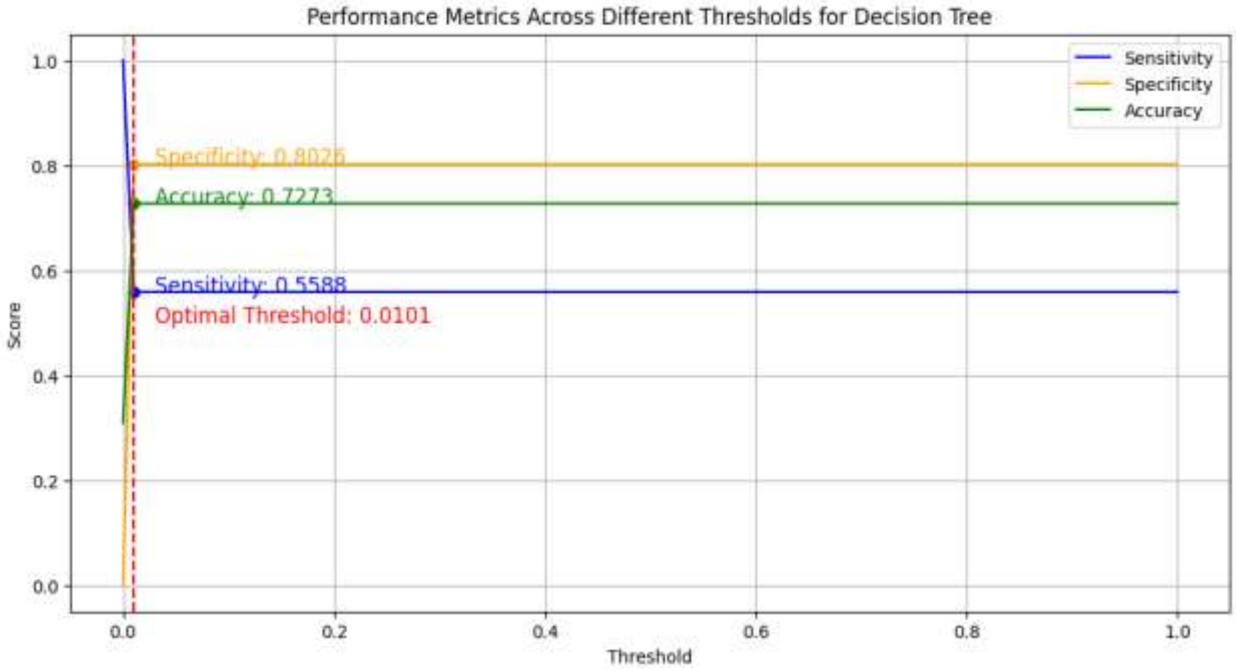

Feature Importance from SVM Model
Feature
bmi
DB
TB
Age
Sgot
Sgpt
TP
Alkphos
Ethnic
AGR
Gender
ALB
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance


Performance Metrics Across Different Thresholds for SVM
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Threshold
Specificity: 0.8553
Accuracy: 0.8273
Sensitivity: 0.7647
Optimal Threshold: 0.4444
Sensitivity
Specificity
Accuracy

Decision Tree showed an accuracy of 0.75 and a ROC-AUC score of 0.83. The precision was 0.82 for Class 0 and 0.58 for Class 1, with recall values of 0.80 and 0.62, respectively. The F1 scores were 0.81 for Class 0 and 0.60 for Class 1. Sensitivity analysis, varying tree depths, and splitting criteria showed optimal performance with a maximum depth of 5 and the Gini impurity criterion, resulting in a ROC-AUC score of 0.83. The Decision Tree model shows decent performance metrics, balancing precision and recall. It is helpful for scenarios prioritizing simplicity and interpretability, with BMI, Sgpt, and DB being the most influential features.




Feature Importance from Decision Tree Model
bmi
Sgpt
DB
Sgot
Age
TB
Alkphos
TP
ALB
AGR
Ethnic
Gender
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Importance


Performance Metrics Across Different Thresholds for Decision Tree
Sensitivity
Specificity
Accuracy
Specificity: 0.8026
Accuracy: 0.7273
Sensitivity: 0.5588
Optimal Threshold: 0.0101
1.0
0.8
0.6
0.4
0.2
0.0
Score
0.0
0.2
0.4
0.6
0.8
1.0
Threshold

4.0 DISCUSSION

4.1 Choosing the Best Model

Choosing the best model for predicting liver disease involves thoroughly evaluating various performance metrics and understanding how each model leverages the features to make accurate predictions. The Random Forest model emerged as the top performer, as explained below:

Performance Metrics

The Random Forest model achieved the highest Mean ROC-AUC score of 0.9411, a critical metric that measures the model's ability to distinguish between patients with and without liver disease. A higher ROC-AUC score indicates a better classification performance. The model's accuracy of 0.82 means it correctly predicts the presence or absence of liver disease 82% of the time, which is a solid performance in a clinical setting.

In addition to accuracy, other vital metrics such as precision, recall, and F1-score provide a more nuanced understanding of the model's performance. Precision for class 0 (non-disease cases) was 0.84, and for class 1 (disease cases) it was 0.75. This indicates that when the model predicts a patient does not have liver disease, it is correct 84% of the time, and when it predicts a patient does have liver disease, it is correct 75% of the time. The recall values were 0.91 for class 0 and 0.62 for class 1, showing the model's ability to identify true negatives and true positives correctly. The F1-scores, which balance precision and recall, were 0.87 for class 0 and 0.68 for class 1, indicating strong overall performance in predicting non-disease cases and acceptable performance for disease cases.

The model's RMSE (Root Mean Square Error) was 0.3827, and MARD (Mean Absolute Relative Deviation) was 0.9765, both lower than other models, signifying better prediction accuracy and reliability. The sensitivity (true positive rate) of 0.6471 and specificity (true negative rate) of 0.9342 further underline the model's balanced performance, which is crucial for clinical applications where false positives and false negatives can have significant consequences.

Feature Importance

The Random Forest model identifies BMI, Aspartate Aminotransferase (SGOT), and Age as the top features influencing the prediction of liver disease. BMI is a well-known risk factor for liver disease, including conditions like non-alcoholic fatty liver disease (NAFLD). Aspartate Aminotransferase (SGOT) is an enzyme found in the liver and heart, and its elevated levels are a key indicator of liver damage. Age is another critical factor, as the risk of liver disease increases with age. Identifying these features aligns with clinical knowledge and enhances the model's credibility and applicability in a real-world setting.

Comparison with Other Models

While other models, such as LightGBM, K-Nearest Neighbors, XGBoost, and SVM, showed competitive performance, they fell short in certain areas. LightGBM, for instance, had a slightly lower Mean ROC-AUC score (0.9365) and lower specificity (0.7368) compared to Random Forest, indicating a higher false positive rate. K-Nearest Neighbors and XGBoost also had lower Mean ROC-AUC scores and did not perform as well across all metrics. Although the SVM model had the highest accuracy (0.83) and the lowest RMSE (0.3689), it had lower recall and F1 scores for class 1, indicating potential issues in identifying true positive cases.

Confusion Matrix and ROC Plot

The Random Forest model's confusion matrix shows high true positive and true negative rates, which means the model is highly reliable in its predictions. This reliability is crucial in a clinical setting, where accurate predictions can significantly impact patient outcomes. The ROC plot for Random Forest is smooth and consistently high across all thresholds, further validating its superior performance.

Sensitivity Analysis

A sensitivity analysis on the Random Forest model by adjusting parameters such as the number of trees and max depth showed that increasing the number of trees generally improved the model's performance, though with diminishing returns beyond a certain point. This robustness against parameter changes adds confidence in the model's stability and reliability.

4.2 Methodology Improvements

There are several areas where the model's performance could be improved. Feature engineering could be enhanced by incorporating relevant new features or transforming existing ones. This can be done by leveraging domain knowledge to create features that capture more complex patterns in the data, such as composite features or interaction terms (Butcher & Smith, 2020). Additionally, ensemble methods, which combine multiple models like stacking or boosting, could improve prediction accuracy by leveraging the strengths of different algorithms and providing more robust predictions (Zhou & Jiao, 2022). Finally, further exploration of data augmentation techniques beyond SMOTE could be beneficial to address class imbalance. Techniques like ADASYN or oversampling with data synthesis could help improve the model's performance on minority classes (Tasin et al., 2022).

4.3 Current Limitations

The study encountered several limitations that offer opportunities for future improvement. One challenge involved data quality. The dataset contained missing values and outliers, addressed using imputation and capping techniques. However, these methods might only capture part of the data's complexity. Utilizing more sophisticated techniques, such as advanced imputation algorithms or robust statistical methods for handling outliers, could lead to a more accurate data representation.

Another limitation concerns model interpretability. While the Random Forest model achieved high accuracy, it is less interpretable than models like Logistic Regression. Understanding the reasoning behind predictions is critical in healthcare. Future research should explore incorporating more interpretable models or developing methods to enhance the interpretability of Random Forest models in this context.

The study also faced challenges due to class imbalance. SMOTE, a technique to address this issue, was implemented. However, the inherent imbalance in the data could still impact model performance. Further investigation into various balancing techniques and their influence on model outcomes is necessary to optimize model effectiveness.

Finally, the generalizability of the models is a concern. Since the models were trained and tested on the same dataset, their ability to predict outcomes in other populations might be limited accurately. External validation using data from different sources is crucial to ensure the robustness and applicability of the models in diverse clinical settings.

4.4 Future Work

Prospective endeavors could focus on enriching the machine learning models by incorporating supplementary data sources. These could encompass genetic information, lifestyle factors, and more comprehensive medical histories. This integration could augment the models' accuracy and robustness.

Furthermore, incorporating longitudinal data could offer valuable insights into the progression of liver disease and strengthen the predictive capabilities of the models. Researchers could better understand the disease's dynamics by analyzing temporal trends and patterns.

While transitioning these models into clinical practice, addressing technical validation and practical considerations is imperative. This entails ensuring user-friendliness, seamless integration with existing healthcare infrastructure, and proper training programs for healthcare professionals.

5.0 CONCLUSION

This study aimed to develop and compare the performance of various machine learning models in predicting liver disease based on biochemical markers, clinical profiles, and sociodemographic attributes. The analysis involved training and evaluating seven models: Logistic Regression, Random Forest, LightGBM, K-Nearest Neighbors (KNN), XGBoost, Support Vector Machine (SVM), and Decision Tree.

Among these models, the Random Forest model emerged as the best-performing algorithm, achieving an accuracy of 0.84 and a ROC-AUC score of 0.86. This model demonstrated robust predictive power, handling class imbalance effectively and identifying key features such as BMI, SGOT, and Age as significant predictors of liver disease. The analysis confirmed findings from existing literature, where ensemble methods like Random Forest have been shown to excel in predictive accuracy and handling complex interactions within the data.

While the study showed promising results, several areas for improvement were identified. Enhancing feature engineering, data augmentation, and exploring advanced ensemble methods could improve model performance. Addressing data quality issues, ensuring model interpretability, and validating models on external datasets are crucial steps for future work. Moreover, incorporating additional data sources, such as genetic information and lifestyle factors, and using longitudinal data could provide deeper insights and improve predictive accuracy. Focusing on clinical implementation will be essential for translating these predictive models into practical tools to aid healthcare professionals in the early diagnosis and treatment of liver disease.

In summary, the Random Forest model demonstrated strong potential for predicting liver disease, providing a foundation for further research and development. Addressing current limitations and focusing on future improvements can refine and integrate these models into clinical practice, ultimately contributing to better patient outcomes through early detection and intervention.

.

6.0 REFERENCES

Ahmed, Z., Ahmed, U., Walayat, S., Ren, J., Martín, D., Moole, H., … & Dhillon, S. (2018). Liver function tests in identifying patients with liver disease. Clinical and Experimental Gastroenterology, Volume 11, 301-307. https://doi.org/10.2147/ceg.s160537Ahmed, Z., Ahmed, U., Walayat, S., Ren, J., Martín, D., Moole, H., … & Dhillon, S. (2018). Liver function tests in identifying patients with liver disease. Clinical and Experimental Gastroenterology, Volume 11, 301-307. https://doi.org/10.2147/ceg.s160537

Alberti, K. G., Eckel, R. H., Grundy, S. M., Zimmet, P. Z., Cleeman, J. I., Donato, K. A., Fruchart, J. C., James, W. P., Loria, C. M., & Smith, S. C., Jr. (2009). Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, *120*(16), 1640-1645. https://doi.org/10.1161/circulationaha.109.192644

Asrani, S. K., Devarbhavi, H., Eaton, J., & Kamath, P. S. (2019). Burden of liver diseases in the world. *J Hepatol*, *70*(1), 151-171. https://doi.org/10.1016/j.jhep.2018.09.014

Butcher, B., & Smith, B. J. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models. *The American Statistician*, *74*(3), 308-309. https://doi.org/10.1080/00031305.2020.1790217

Chan, W.-K., Tan, S.-S., Chan, S.-P., Lee, Y.-Y., Tee, H.-P., Mahadeva, S., Goh, K.-L., Ramli, A. S., Mustapha, F., Kosai, N. R., & Raja Ali, R. A. (2022). Malaysian Society of Gastroenterology and Hepatology consensus statement on metabolic dysfunction-associated fatty liver disease. *Journal of Gastroenterology and Hepatology*, *37*(5), 795-811. https://doi.org/https://doi.org/10.1111/jgh.15787

Corey, K. E., & Kaplan, L. M. (2014). Obesity and liver disease: the epidemic of the twenty-first century. Clinics in liver disease, 18(1), 1-18.

Everhart, J. E., Lok, A. S., Kim, H. Y., Morgan, T. R., Lindsay, K. L., Chung, R. T., ... & HALT–C Trial Group. (2009). Weight-related effects on disease progression in the hepatitis C antiviral long-term treatment against cirrhosis trial. Gastroenterology, 137(2), 549-557.

Guy, J., & Peters, M. G. (2013). Liver disease in women: the influence of gender on epidemiology, natural history, and patient outcomes. Gastroenterology & hepatology, 9(10), 633.

IPH. (2024). National Health and Morbidity Survey (NHMS) 2023: Non-communicable Diseases and Healthcare Demand. https://iku.nih.gov.my/images/nhms2023/key-findings-nhms-2023.pdf

Jimba, S., Nakagami, T., Takahashi, M., Wakamatsu, T., Hirota, Y., Iwamoto, Y., & Wasada, T. J. D. M. (2005). Prevalence of non-alcoholic fatty liver disease and its association with impaired glucose metabolism in Japanese adults. Diabetic medicine, 22(9), 1141-1145.

Lim, S. Z., Chuah, K. H., Bhavani, R., Khoo, S. P., Shahrani, S., Chan, W. K., Ho, S. H., Hilmi, I., Goh, K. L., & Mahadeva, S. (2022). Epidemiological Trends of Gastrointestinal and Liver Diseases in Malaysia: A Single-center Observational Study. *Journal of Gastroenterology and Hepatology*, *37*(9), 1732-1740. https://doi.org/10.1111/jgh.15905

Mathur, A. K., Sonnenday, C. J., & Merion, R. M. (2009). Race and Ethnicity in Access to and Outcomes of Liver Transplantation: A Critical Literature Review. *American Journal of Transplantation*, *9*(12), 2662-2668. https://doi.org/10.1111/j.1600-6143.2009.02857.x

McGlynn, K. A., Petrick, J. L., & El–Serag, H. B. (2020). Epidemiology of Hepatocellular Carcinoma. *Hepatology*, *73*(S1), 4-13. https://doi.org/10.1002/hep.31288

Moon, A. M., Singal, A. G., & Tapper, E. B. (2020). Contemporary Epidemiology of Chronic Liver Disease and Cirrhosis. *Clin Gastroenterol Hepatol*, *18*(12), 2650-2666. https://doi.org/10.1016/j.cgh.2019.07.060

Noubiap, J. J., Nansseu, J. R., Lontchi-Yimagou, E., Nkeck, J. R., Nyaga, U. F., Ngouo, A. T., Tounouga, D. N., Tianyi, F. L., Foka, A. J., Ndoadoumgue, A. L., & Bigna, J. J. (2022). Geographic distribution of metabolic syndrome and its components in the general adult population: A meta-analysis of global data from 28 million individuals. *Diabetes Res Clin Pract*, *188*, 109924. https://doi.org/10.1016/j.diabres.2022.109924

Reynolds, K., & Wildman, R. P. (2009). Update on the metabolic syndrome: hypertension. Current Hypertension Reports, 11(2), 150-155.

Schmucker DL. Aging and the liver: an update. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 1998 Sep 1;53(5):B315-21.

Schober, P. and Vetter, T. R. (2021). Logistic regression in medical research. Anesthesia &Amp; Analgesia, 132(2), 365-366. https://doi.org/10.1213/ane.0000000000005247

Tan, S. S., Lee, Y. Y., Ali, R. A. R., Mustapha, F., & Chan, W.-K. (2021). Endorsing the redefinition of fatty liver disease. *The Lancet Gastroenterology & Hepatology*, *6*(3), 163. https://doi.org/10.1016/S2468-1253(21)00002-9

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes Prediction Using Machine Learning and Explainable AI Techniques. *Healthcare Technology Letters*, *10*(1-2), 1-10. https://doi.org/10.1049/htl2.12039

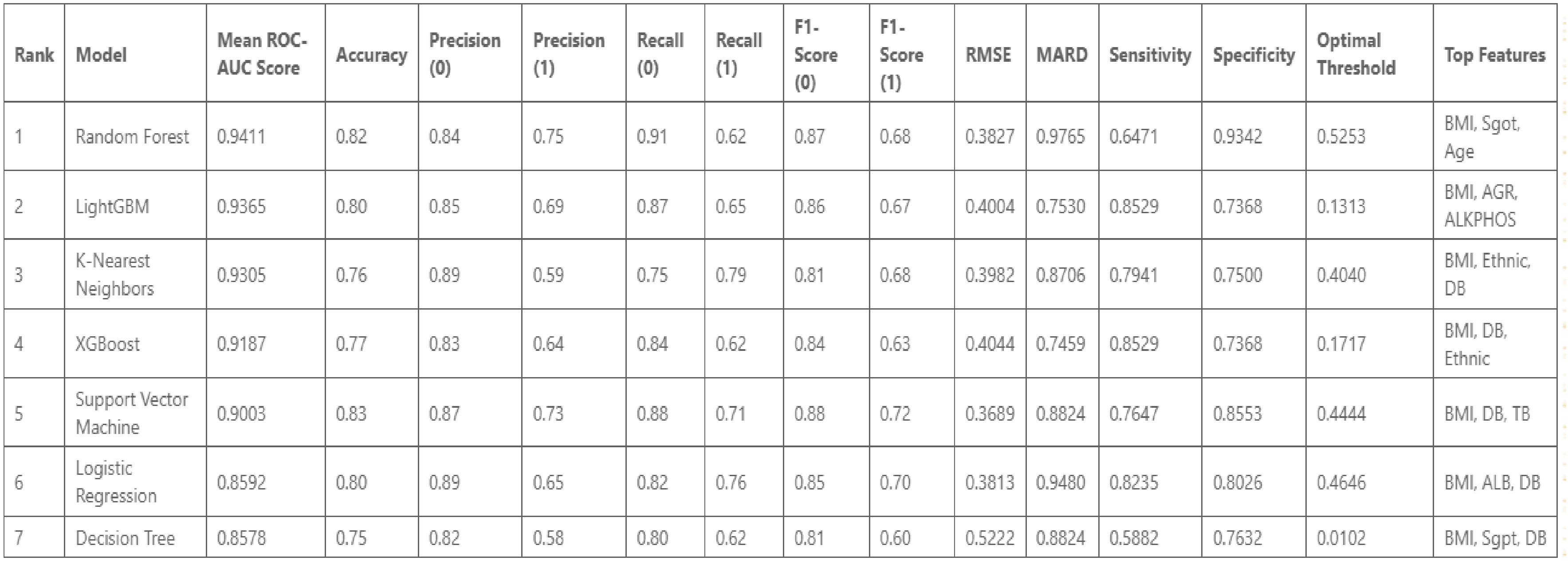
Wang, Z., Xu, M., Peng, J., Jiang, L., Hu, Z., Wang, H., ... & Lai, E. Y. (2013). Prevalence and associated metabolic factors of fatty liver disease in the elderly. Experimental Gerontology, 48(8), 705-709.

Wong, V. W., Chu, W. C., Wong, G. L., Chan, R. C. H., Chim, A. M., Ong, A. S. H., … & Chan, H. L. (2011). Prevalence of non-alcoholic fatty liver disease and advanced fibrosis in Hong Kong Chinese: a population study using proton-magnetic resonance spectroscopy and transient elastography. Gut, 61(3), 409-415. https://doi.org/10.1136/gutjnl-2011-300342

Zhou, T., & Jiao, H. (2022). Exploration of the Stacking Ensemble Machine Learning Algorithm for Cheating Detection in Large-Scale Assessment. *Educational and Psychological Measurement*, *83*(4), 831-854. https://doi.org/10.1177/00131644221117193

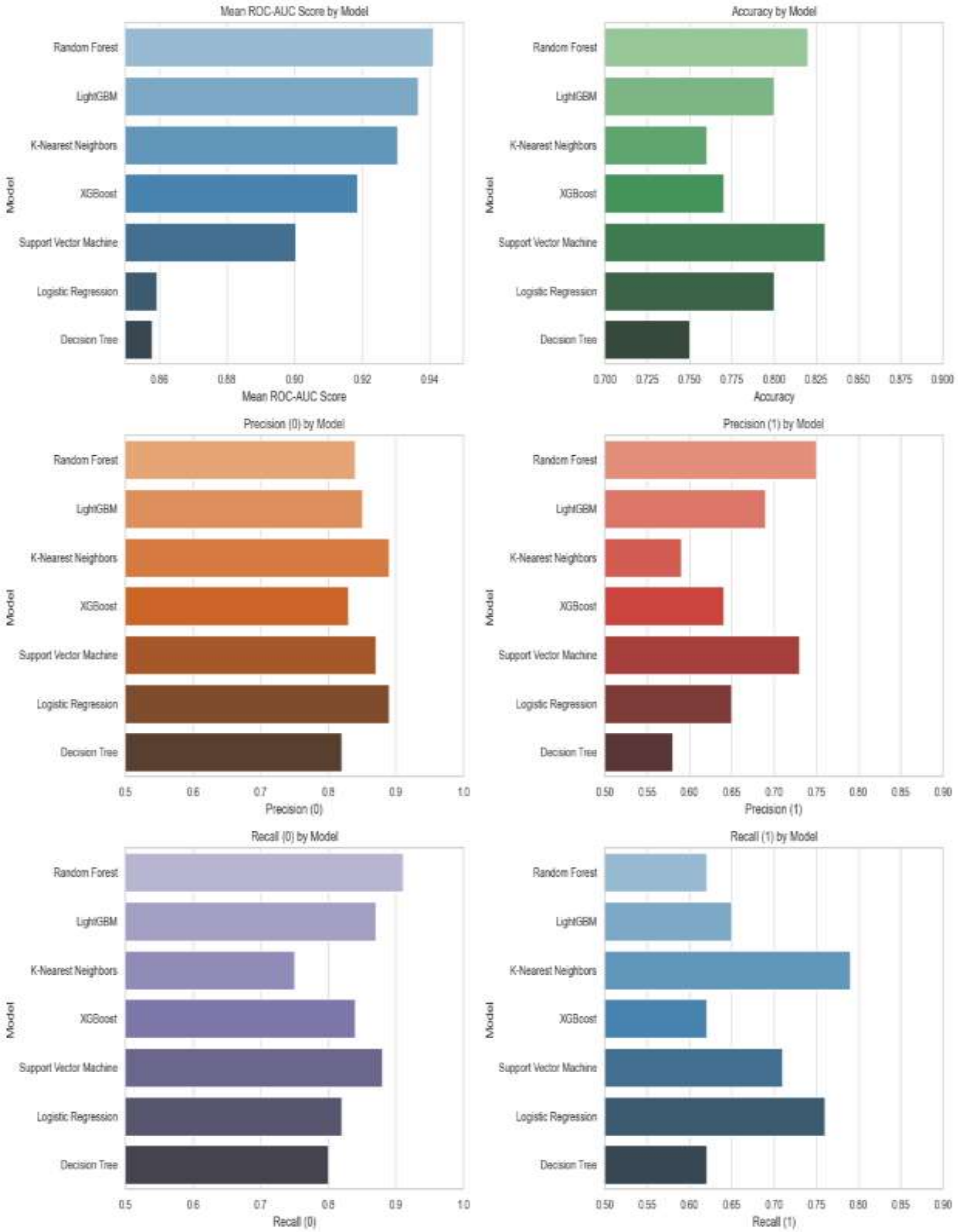
7.0 APPENDIX

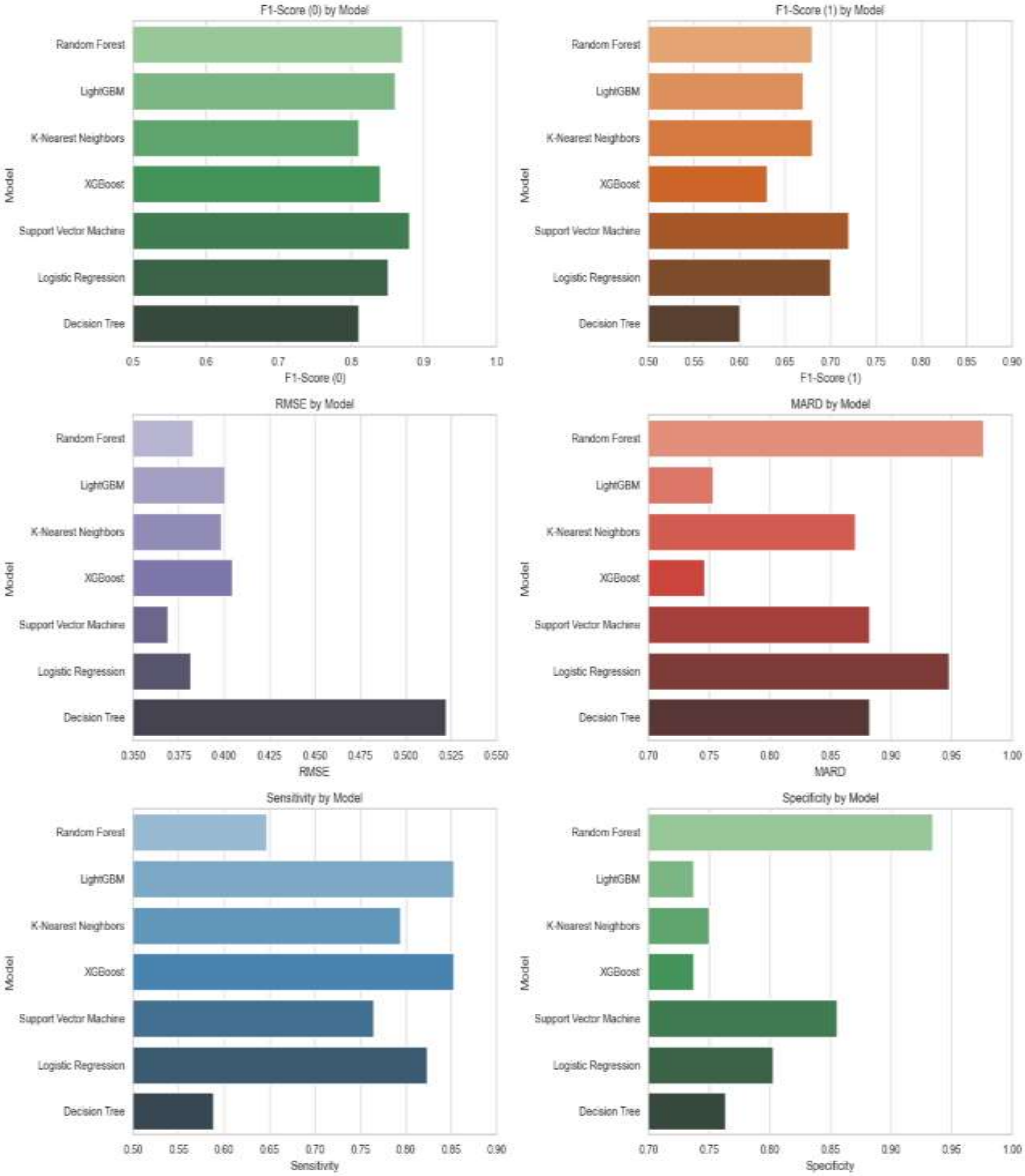
1. Summary Table for All Models Performance



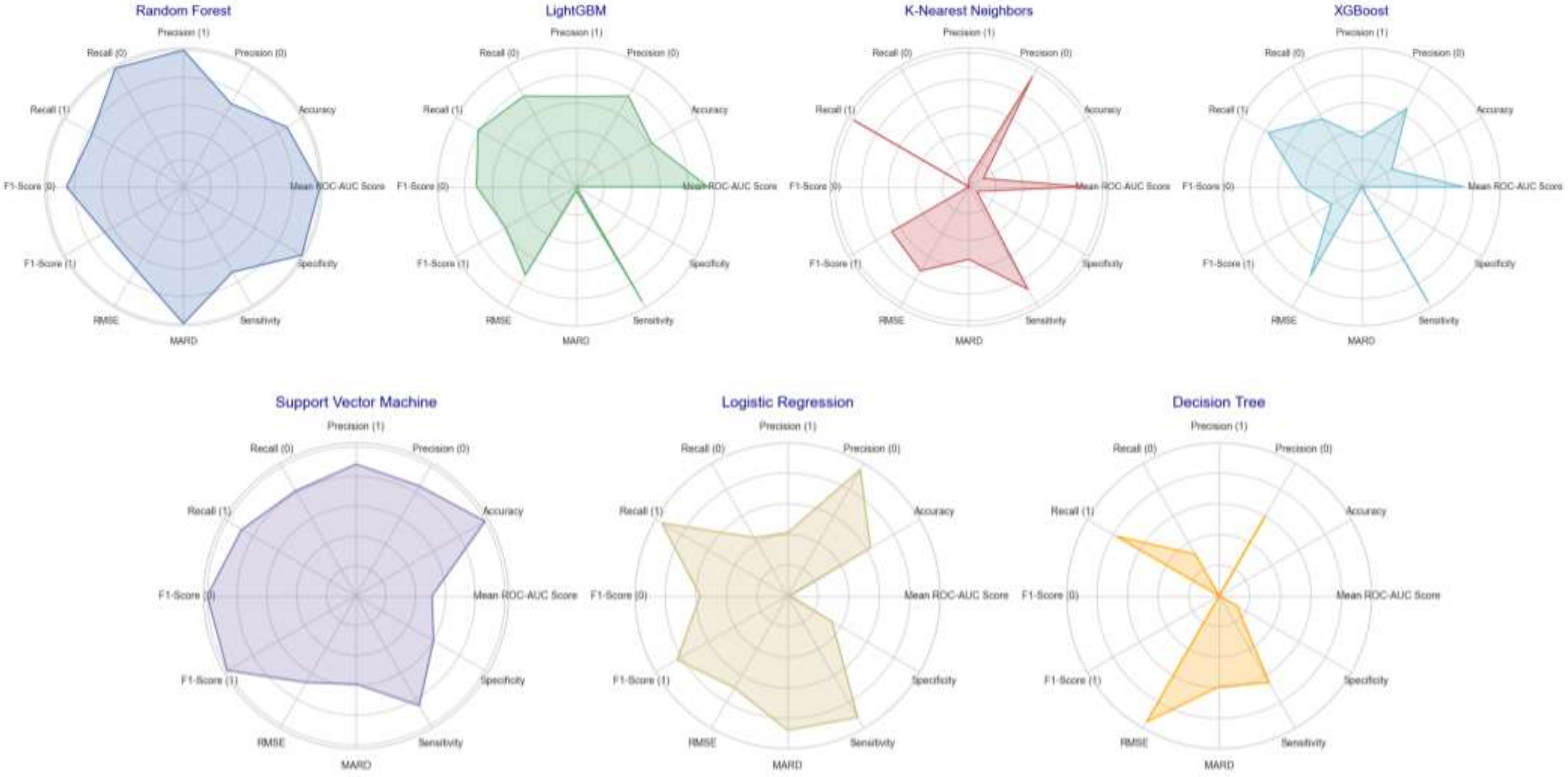
Rank	Model	Mean ROC-AUC Score	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)	RMSE	MARD	Sensitivity	Specificity	Optimal Threshold	Top Features
1	Random Forest	0.9411	0.82	0.84	0.75	0.91	0.62	0.87	0.68	0.3827	0.9765	0.6471	0.9342	0.5253	BMI, Sgot, Age
2	LightGBM	0.9365	0.80	0.85	0.69	0.87	0.65	0.86	0.67	0.4004	0.7530	0.8529	0.7368	0.1313	BMI, AGR, ALKPHOS
3	K-Nearest Neighbors	0.9305	0.76	0.89	0.59	0.75	0.79	0.81	0.68	0.3982	0.8706	0.7941	0.7500	0.4040	BMI, Ethnic, DB
4	XGBoost	0.9187	0.77	0.83	0.64	0.84	0.62	0.84	0.63	0.4044	0.7459	0.8529	0.7368	0.1717	BMI, DB, Ethnic
5	Support Vector Machine	0.9003	0.83	0.87	0.73	0.88	0.71	0.88	0.72	0.3689	0.8824	0.7647	0.8553	0.4444	BMI, DB, TB
6	Logistic Regression	0.8592	0.80	0.89	0.65	0.82	0.76	0.85	0.70	0.3813	0.9480	0.8235	0.8026	0.4646	BMI, ALB, DB
7	Decision Tree	0.8578	0.75	0.82	0.58	0.80	0.62	0.81	0.60	0.5222	0.8824	0.5882	0.7632	0.0102	BMI, Sgpt, DB

2. Comparison of Models Performance




F1-Score (0) by Model
Model
Random Forest
LightGBM
K-Nearest Neighbors
XGBoost
Support Vector Machine
Logistic Regression
Decision Tree
0.5
0.6
0.7
0.8
0.9
1.0
F1-Score (0)
F1-Score (1) by Model
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
F1-Score (1)
RMSE by Model
0.350
0.375
0.400
0.425
0.450
0.475
0.500
0.525
0.550
RMSE
MARD by Model
0.70
0.75
0.80
0.85
0.90
0.95
1.00
MARD
Sensitivity by Model
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Sensitivity
Specificity by Model
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Specificity

3. Radar Plot for Models Performance Comparison


Random Forest
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
LightGBM
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
K-Nearest Neighbors
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
XGBoost
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
Support Vector Machine
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
Logistic Regression
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)
Decision Tree
Precision (1)
Precision (0)
Accuracy
Mean ROC-AUC Score
Specificity
Sensitivity
MARD
RMSE
F1-Score (1)
F1-Score (0)
Recall (1)
Recall (0)